N° d'ordre : 2874

THÈSE

présentée

devant l'université de Rennes 1

pour obtenir

le grade de : Docteur de l'université de Rennes 1
Mention Traitement du Signal

par

Elie Laurent BENAROYA

Équipe d'accueil : METISS
École doctorale : MATISSE
Composante universitaire : IFSIC/IRISA

Titre de la thèse :

Séparation de plusieurs sources sonores avec un seul microphone

Soutenue le 26 Juin 2003 devant la commission d'examen

M. :	Bernard	Delyon	Président
M. :	Eric	Moulines	Rapporteur
M. :	Ali	Mohammad-Djafari	Rapporteur
M. :	Jean-Jacques	Bellanger	Examinateur
M. :	Eric	Le Carpentier	Examinateur
M. :	Frédéric	Bimbot	Directeur de thèse

A esprit libre,
univers libre.

Koan

A Joachim et à mes parents.

Remerciements

Je tiens en premier lieu à remercier Frédéric Bimbot, Chargé de Recherche et Responsable Scientifique de l'équipe METISS au sein du laboratoire IRISA, pour avoir encadré mes travaux pendant toute cette thèse. Je tiens à souligner la qualité de son encadrement scientifique ainsi que son soutien moral au quotidien qui ont permis la réalisation de cette thèse dans de bonnes conditions.

Je suis tout particulièrement reconnaissant aux rapporteurs de cette thèse, Eric Moulines, Professeur au Département du Traitement du Signal et des Images de l'Ecole Nationale Supérieure des Télécommunications de Paris et Ali Mohammad-Djafari, Directeur de Recherche au Laboratoire des Signaux et Systèmes de l'Ecole Supérieure d'Electricité de Paris, pour l'attention qu'ils ont porté à ces travaux. Je remercie Bernard Delyon, Professeur à l'Université de Rennes I, qui a bien voulu présider le jury de soutenance de cette thèse, ainsi que Jean-Jacques Bellanger, Maître de Conférence à l'Université de Rennes I, et Eric Le Carpentier, Maître de conférence à l'Institut de Recherche en Communications et en Cybernétique de Nantes, d'avoir bien voulu examiner ce travail.

Je tiens à témoigner ma reconnaissance envers Lorcan Mc Donagh avec qui j'ai eu la joie de partager mon bureau, pour sa disponibilité, tant scientifique qu'amicale ainsi que pour sa bonne humeur quotidienne. Je remercie tous les membres passés ou présents de l'équipe METISS, Mathieu Ben, Mickaël Betser, Rémi Gribonval, Guillaume Gravier, Mouhamadou Seck et Fabienne Porée qui ont contribué au cadre agréable de travail tout au long de cette thèse.

Je remercie particulièrement Raphaël Blouet pour nos toutes discussions, scientifiques ou pas, et pour son soutien sans faille.

Résumé

Le problème de la séparation de sources sonores dans des conditions quasi-réelles, c'est-à-dire avec peu de microphones, suscite un intérêt croissant de la part de la communauté du traitement du signal. Dans ce cadre, le problème de séparation de sources avec un capteur unique a été peu étudié et est en émergence.

L'objet de cette thèse est l'étude de la séparation de sources sonores avec un seul capteur, dans le domaine temps-fréquence. Deux méthodes permettant de résoudre partiellement ce problème sont exposées d'un point de vue théorique et illustrées sur des exemples réels. Ces méthodes permettent d'étendre les techniques classiques de filtrage fréquentiel (filtrage de Wiener) à des signaux non-stationnaires et en particulier dans le cadre de signaux stationnaires à court terme. La première méthode est une extension du filtrage de Wiener à des modèles de mélange de gaussiennes pour les sources. La seconde méthode est fondée sur une décomposition non négative du spectre du mélange sur un dictionnaire de formes spectrales caractéristiques des deux sources.

Outre des contributions au niveau de la formalisation et de l'algorithmique, un des apports notables des travaux présentés dans cette thèse concerne l'utilisation de modèles à facteur d'amplitude et leur application pour la modélisation de sources stationnaires à court terme.

D'autre part, une évaluation comparative des algorithmes proposés dans cette thèse est fournie sur un ensemble de signaux sonores réels et les modalités d'implémentation des algorithmes sont traités.

Cette thèse est donc une exploration des possibilités d'extensions du filtrage de Wiener pour la séparation de sources avec un seul capteur et c'est aussi, nous l'espérons, un point de départ pour de nouveaux développements tant théoriques que pratiques.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'objectif de ce travail est d'étudier les possibilités et les modalités d'une séparation de sources sonores dans le cas où un seul capteur ou microphone est disponible. L'étude de ce type de problème est très récente : la bibliographie est donc succincte. Néanmoins, la séparation de sources en général est un sujet à la mode et très actif. Nous présenterons donc dans les grandes lignes les principales approches pour ce problème, notamment dans le cadre de l'Analyse en Composantes Indépendantes (ACI).

Les premières tentatives de séparation de sources avec un capteur unique sont apparues dans le cadre de l'approche CASA (Computational Auditory Scene Analysis). Cette approche a pour but de reproduire ou d'approcher sous forme d'algorithmes la manière dont nous percevons les sons et il est donc naturel dans ce cadre de s'intéresser à la séparation de sources sonores à partir d'un enregistrement unique, puisqu'une oreille entraînée est capable de distinguer différents instruments de musique et de se focaliser sur l'un d'eux, dans le cas d'un enregistrement polyphonique ou composite.

Nous ne présenterons pas dans le détail l'approche CASA et les différents algorithmes qui en résultent, car cette approche nous apparaît plus comme un point de départ qu'un centre incontournable dans le cadre du problème que nous nous posons. Cela se justifie d'autant plus que la séparation de source n'est pas le coeur de l'approche CASA, mais seulement un domaine d'application.

Néanmoins, nous étudierons dans le contexte de l'état de l'art les propositions de Sam Roweis dans son article *One Microphone Source Separation* [Row00], 2000. Cet auteur a tenté de rapprocher la philosophie CASA et le traitement du signal dans le but de décrire un système capable de séparer des sources sonores avec un seul capteur. Comme nous le verrons par la suite, le système que présente Roweis a de nombreux points communs avec une des approches que nous présenterons dans ce document : la généralisation du filtre de Wiener à des modèles multi-gaussiens. Nous espérons éclairer la tentative de Roweis d'un jour nouveau à travers les développements théoriques que nous présenterons dans ce document. Notons

aussi que dans le cadre du traitement de la parole en milieu bruité, il existe des méthodes qui combinent modèles à états cachés (Modèle de Markov cachés à densité conditionnelle gaussienne) et débruitage (tel que les travaux de Y. Ephraïm ou R. Varga et R.K. Moore). Les travaux présentés ici sont dans cette même lignée et en sont, peut-être, une généralisation pour la séparation de signaux audio.

Les principaux résultats présentés dans cette thèse concernent deux méthodes de séparation de sources sonores dans le cas d'un capteur unique. Ces deux méthodes ont en commun une phase d'apprentissage, dans laquelle des formes caractéristiques du spectres ou DSP (Densité Spectrale de Puissance) sont estimées sur des exemples de sources séparées. La première méthode est fondée sur une généralisation du filtrage de Wiener à des lois *a priori* non gaussiennes, pour être précis des mélanges de gaussiennes (MMG). Cette généralisation permet de traiter des signaux non-stationnaires et donne lieu à un schéma de filtrage adaptatif. La seconde méthode est fondée sur une représentation parcimonieuse et non-négative du spectre du signal composite sur l'ensemble des DSP disponibles. Nous dérivons dans ce cas une formule analogue à celle de Wiener. Nous verrons que les performances des deux algorithmes sont comparables, avec des temps de calcul plus faibles dans le cas de la deuxième méthode (la complexité algorithmique est moindre).

L'un des apports les plus marquants de cette thèse est l'utilisation et la généralisation de modèles gaussiens à facteur d'amplitude (Gaussian Scaled Mixture Models, en anglais), pour la modélisation de signaux localement stationnaires. Ces modèles sont appliqués dans cette thèse aux extensions du filtrage de Wiener et semblent représenter une amélioration importante pour l'application visée, c'est-à-dire la séparation de sources audio avec un seul microphone.

Enfin, remarquons que dans le cadre de l'analyse en composantes indépendantes (ACI), on parle souvent de méthodes dites *aveugles* de séparation de sources. Cette terminologie correspond à un problème dit de données incomplètes et il est assez naturel d'utiliser des algorithmes de type EM (Expectation-Maximization) dans ce cas. Nous verrons le rôle de cet algorithme standard dans le cadre des mélanges sous déterminés (plus de sources que de capteurs), dans la partie bibliographie.

Dans le cadre de l'ACI, le formalisme de données incomplètes correspond à l'absence de connaissances sur la matrice de mélange, pour l'estimation des sources. Cette matrice correspond à la manière dont les sources s'additionnent pour former les différents mélanges observés.

Dans le cas d'un capteur unique, le terme de *séparation aveugle* ne s'applique plus vraiment, car toutes les tentatives actuelles pour résoudre ce problème s'accordent sur la nécessité

d'une phase préliminaire d'apprentissage sur des sources séparées. Ces exemples de sources séparées, connus à l'avance, ne sont pas nécessairement strictement identiques à ceux contenus dans le mélange, mais permettent d'extraire des caractéristiques propres à chacune des sources (les DSP par exemple). Néanmoins, il demeure des paramètres non observés et non connus à l'avance tels que les séquences d'indexes des gaussiennes dans le cas de modèles de mélange de gaussiennes ou les facteurs d'amplitudes qui 'modulent' les DSP. Ainsi, les algorithmes liés au problème de données incomplètes peuvent trouver leur place dans le cadre de ce que nous allons présenter.

En fait, le cadre très général de l'analyse en composantes indépendantes (ACI) ne permet pas de traiter le cas du capteur unique, car ce formalisme repose sur l'existence et l'estimation d'une matrice de mélange, qui dans le cas d'un capteur unique est complètement dégénérée et n'apporte pas d'information utile. Néanmoins, ce que nous présentons dans cette thèse partage avec l'ACI un formalisme commun : la théorie bayésienne. Celle-ci permet dans un cadre simple et cohérent de tenir compte des informations *a priori* sur le phénomène observé.

Plan des travaux

Ce travail se décompose en cinq parties :

- Présentation du sujet
- Approche probabiliste bayésienne
- Modèles à facteurs d'amplitude
- Expérimentation et évaluation
- Conclusions et perspectives

Présentation

Le premier chapitre de cette partie concerne l'état de l'art de la séparation de sources. Nous décrivons l'approche par Analyse en Composantes Indépendantes (ACI) de la séparation de sources dans les cas déterminé et sous-déterminé (moins de microphones que de sources). Ensuite, nous discutons différentes approches récentes de la séparation de sources avec un seul microphone.

Le second chapitre est consacré à la propriété de stationnarité locale qui est en général vraie pour les signaux audio. Cette notion est définie théoriquement et nous en déduisons les hypothèses de travail que nous utiliserons dans toute la suite de la thèse.

Le troisième chapitre traite de la problématique de l'évaluation des algorithmes de séparation de sources, afin de définir quelles doivent être les propriétés désirées des critères d'évaluation

en vue de leur utilisation lors de l'expérimentation.

Approche probabiliste bayésienne

Dans le premier chapitre, nous rappelons les bases de la théorie de l'estimation bayésienne. En effet, dans toute la thèse, nous utiliserons des modèles probabilistes pour les sources dans le cadre du formalisme bayésien. Ce formalisme permet d'intégrer dans un cadre théorique cohérent les informations *a priori* sur les sources, c'est-à-dire apprises dans une phase préalable.

Le second chapitre présente une application classique de la théorie bayésienne : le filtre de Wiener. Ce chapitre servira de base aux extensions ultérieures.

Dans le troisième chapitre, nous décrivons notre approche de la séparation de source avec un seul microphone à l'aide du filtre de Wiener étendu à des Modèles de Mélanges de Gausssiennes (MMG) comme densités *a priori* des sources.

Dans le quatrième chapitre, nous traitons de la modélisation du logarithme du spectre de puissance des signaux dans le cadre de modèles de mélange de gaussiennes.

Modèles à facteur d'amplitude

Après un premier chapitre introductif à propos de ces modèles, nous développons dans le second chapitre deux manières de combiner les modèles gaussiens à facteur d'amplitude afin d'obtenir un modèle réaliste de source sonore. Cela donnera donc lieu à deux algorithmes différents d'estimation des sources : l'un fondé sur les modèles de mélange de gaussiennes à facteur d'amplitude (MMGA) et l'autre sur les dictionnaires de DSP.

Le troisième chapitre aborde la question délicate de l'estimation des facteurs d'amplitude lors de la séparation (démixage).

Enfin, le chapitre quatre concerne l'apprentissage de dictionnaires de DSP.

Expérimentation et évaluation

Le premier chapitre de cette partie est consacré aux critères d'évaluation que nous avons définis au cours de cette thèse dans le cadre du groupe de travail "Ressources pour la séparation de signaux audiophoniques" (Action Jeunes Chercheurs du GdR ISIS).

Dans le second chapitre, nous expliquons les aspects pratiques d'implémentation des algorithmes définis dans la partie précédente.

Enfin, nous terminons cette partie par une évaluation comparative des méthodes de séparation de sources proposées dans cette thèse sur des signaux sonores réels.

Conclusion et perspectives

Nous consacrerons un chapitre sur des questions non résolues dans le cadre de ce travail, notamment en ce qui concerne la modélisation de la phase des signaux sonores et les raffinements possibles des modèles de sources sonores.

Nous terminerons ce travail par une conclusion ainsi qu'une ébauche des différentes perspectives qui s'ouvrent à nos recherches.

Première partie

Présentation

Cette partie est introductive et est composée de trois chapitres indépendants :

Le premier chapitre concerne l'état de l'art en séparation de sources et en particulier dans le cas d'un capteur unique. Nous présentons néanmoins des aspects de la séparation de sources dans le cas général en guise d'introduction. Nous distinguons dans ce chapitre trois cas : autant de capteurs que de sources, moins de capteurs que de sources et enfin le cas du capteur unique, qui est le cas étudié dans ce travail.

Le second chapitre concerne les hypothèses que nous utiliserons dans le cadre de la séparation de sources sonores. Il s'agit de la stationnarité locale des signaux audio. Après une présentation théorique de cette propriété, nous posons les hypothèses fondamentales de modélisation des signaux sonores que nous utiliserons pour modéliser les sources.

Le troisième chapitre aborde le problème de l'évaluation du résultat de la séparation de signaux sonores et notamment les différences avec l'évaluation dans le cas surdéterminé où la définition d'une matrice de mélange est possible. Cette introduction à de nouveaux critères d'évaluation trouvera sa justification dans la partie IV qui concerne l'expérimentation sur des signaux réels.

Chapitre 1

État de l'art

Dans l'état de l'art, nous distinguerons trois cas différents pour la séparation de sources en fonction du nombre de capteurs relativement au nombre de sources :

1. autant de capteurs que de sources,
2. moins de capteurs que de sources, mais au moins deux capteurs,
3. un seul capteur (cas traité dans la suite du manuscrit)

L'essor des méthodes à base d'Analyse en Composantes Indépendantes, en particulier depuis l'article *An information-maximization approach to blind separation and blind deconvolution* [BS95] de Bell et Sejnowski en 1995, en font une technique de base incontournable pour la séparation de source. Nous commencerons par présenter, dans le cadre de mélanges linéaires et instantanés, le cas dit " carré ", c'est-à-dire lorsqu'il y a autant de capteurs que de sources.

Nous nous intéresserons ensuite aux cas sous-déterminés, notamment à travers l'approche EM proposée par Bermond et Cardoso [BC99a], [BMC97].

L'étude de l'Analyse en Composantes Indépendantes n'a qu'une valeur d'introduction dans le cadre du problème posé dans cette thèse, mais elle est nécessaire pour établir des repères concernant notre travail et d'autre part en raison de son succès pour certains problèmes de séparations de sources, en pratique.

Enfin, nous aborderons la bibliographie proprement dite, c'est-à-dire relative à la séparation de sources (audio en général) avec un seul capteur. La plupart des articles sur ce sujet sont récents, voire très récents. Cette partie, bien que très succincte, permettra au lecteur de situer ensuite nos travaux par rapport aux différentes approches actuelles de ce problème.

1.1 Les méthodes d'Analyse en Composantes Indépendantes

L'Analyse en Composantes Indépendantes peut être vue comme un prolongement de l'Analyse en Composantes Principales, mais dans laquelle la décorrelation des observations est remplacée par une hypothèse d'indépendance des sources (hypothèse donc plus forte). Nous allons revoir d'abord les fondements de l'Analyse en Composantes Principales, puis développer un aspect de la théorie pour l'ACI.

1.1.1 Rappels d'Analyse en Composantes Principales

On se place dans le cas instantané linéaire et où il y a autant de sources que de capteurs. Les sources sont notées $s_j(t)$ et les observations (mélanges) sont notés $x_i(t)$. On note $A = \{a_{ij}\}$ la matrice de mélange. L'équation du mélange s'écrit alors :

$$x_i(t) = \sum_{j=1}^{n} a_{ij} s_j(t) \, . \tag{1.1}$$

On peut encore noter cette équation sous la forme matricielle suivante :

$$X = AS, \tag{1.2}$$

où l'on a noté, pour le mélange :

$$X = \begin{Bmatrix} x_1(t_1) & \dots & x_1(t_N) \\ \vdots & & \vdots \\ x_n(t_1) & \dots & x_n(t_N) \end{Bmatrix}, \tag{1.3}$$

et de même pour les sources :

$$S = \begin{Bmatrix} s_1(t_1) & \dots & s_1(t_N) \\ \vdots & & \vdots \\ s_n(t_1) & \dots & s_n(t_N) \end{Bmatrix} . \tag{1.4}$$

On définit alors la matrice de covariance empirique des observations $C = \frac{1}{N} X X^T$.

La condition de décorrélation des sources implique que la matrice de covariance des sources soit diagonale. Dans ce cadre, il semble naturel de diagonaliser la matrice de covariance C (en base orthonormale). Ainsi on peut écrire :

$$C = U D U^T, \tag{1.5}$$

où U est une matrice orthogonale, i.e. $UU^T = I$, et D est une matrice diagonale.

Une solution consiste alors à poser $\hat{S} = U^T X$ et on a

$$\frac{1}{N}\hat{S}\hat{S}^T = U^T C U, \tag{1.6}$$

$$= D\,. \tag{1.7}$$

Les sources estimées $\hat{S}$ sont donc décorrélées, mais comme on a posé $A = U$, on impose à la matrice de mélange d'être orthogonale, ce qui n'a en fait aucune raison d'être dans la réalité.

Nous pouvons d'autre part nous imposer la condition de sphéricité suivante

$$\frac{1}{N}\hat{S}\hat{S}^T = I, \tag{1.8}$$

qui est une condition *a priori* plus forte. En fait, les sources ne peuvent être ré-estimées qu'à un facteur d'amplitude près et aux permutations près. En conséquence, cette condition de sphéricité se justifie tout à fait.

La matrice de mélange $A = UD^{1/2}$ est une solution de (1.8), ainsi que toute matrice $\bar{A} = UD^{1/2}V^T$ où V est une matrice orthogonale quelconque. Il y a donc une indétermination à une matrice orthogonale près sur la matrice de mélange, pour cette condition d'indépendance à l'ordre deux. Cela nous conduit à utiliser une condition d'indépendance à un ordre supérieur, notamment à travers la minimisation de l'information mutuelle.

1.1.2 Analyse en Composantes Indépendantes

L'analyse en composantes indépendantes (ACI) se propose d'utiliser des moments d'ordre supérieur à deux dans le but d'obtenir une indépendance statistique des sources et notamment de lever l'indétermination qui existe si l'on se contente de l'ACP. On peut voir l'ACI comme une méthode liée à la théorie de l'information et notamment au concept d'information mutuelle, mais il existe bien d'autres approches pour ce domaine.

Pour illustrer le propos, plaçons-nous de nouveau dans le cas instantané linéaire où il y a autant de sources que de capteurs n. Les équations du mélange s'écrivent (on reprend les notations de la sous-section précédente) :

$$X = AS, \tag{1.9}$$

ou bien

$$S = WX, \tag{1.10}$$

avec $W = A^{-1}$.

On suppose cette fois que les sources sont indépendantes, c'est-à-dire que les densités se factorisent, soit :

$$p(s_1(t), \ldots, s_n(t)) \quad = \quad \prod_{i=1}^{n} p_i(s_i(t)) \, . \tag{1.11}$$

Un moyen d'arriver à cette indépendance statistique ou du moins de s'en approcher, consiste à minimiser l'information mutuelle [CT91], dont l'expression est la suivante :

$$I(S) \quad = \quad \int \log \left[\frac{p(s_1, \ldots, s_n)}{\prod_{j=1}^{n} p_j(s_j)} \right] p(s_1, \ldots, s_n) \, ds_1 \ldots ds_n \, . \tag{1.12}$$

En effet, l'*information mutuelle* est une quantité toujours positive et lorsque $I(S)$ est nulle, alors les sources sont statistiquement indépendantes : c'est en quelque sorte une mesure de l'indépendance statistique des sources. Il est donc naturel de minimiser cette quantité dans le cadre du problème posé par l'ACI. Nous utiliserons simplement un algorithme du gradient, car le calcul de cette quantité (l'information mutuelle) n'est pas simple, alors qu'au niveau de sa dérivée les choses se simplifient naturellement.

Après calculs (voir [LGBS98] par exemple), on obtient une fonction de contraste à minimiser de la forme

$$\left\{ \begin{array}{rcl} C(W) & = & \frac{1}{T} f(S) S^T - I, \\ S & = & WX, \end{array} \right. \tag{1.13}$$

où I est la matrice identité et f est une fonction non linéaire, liée aux densités marginales des sources.

On utilise alors la règle de mise à jour de la matrice W suivante (inverse de la matrice de mélange), en utilisant le gradient naturel [Ama98] :

$$W^{k+1} \quad = \quad W^k - \nu_k \left[\frac{1}{T} f(S^k) {S^k}^T - I \right] W^k. \tag{1.14}$$

Lorsque le contraste s'annule, c'est-à-dire $\frac{1}{T} f(S^\star) {S^\star}^T = I$ avec $S^\star = W^\star X$, un moment d'ordre supérieur (pour des fonctions f_i non linéaires) est diagonal. On a donc $E[f_i(s_i) s_j] = \delta_{i-j}$ où δ est le symbole de Kronecker.

Cela généralise d'une part la condition de décorrélation à l'ordre deux et on devine de plus que la connaissance exacte des fonctions f_i, qui sont liées aux lois des sources séparées, n'est pas forcément nécessaire (voir [AC97] sur ces aspects, d'un point de vue théorique).

Notons qu'il existe d'autres approches de l'ACI, notamment fondées sur la non-gaussiannité des sources. En effet, les observations qui résultent de la superposition de plusieurs sources seront "plus" gaussiennes que les sources elles-mêmes. On peut donc essayer de chercher la

matrice de mélange A qui maximise un critère de non-gaussianité des sources estimées. Parmi ces méthodes, citons les méthodes fondées sur la maximisation de la valeur absolue du kurtosis (moment d'ordre 4) [Car98]. Dans ce cas on parlera encore de fonction de contraste à minimiser.

Notons qu'il existe d'autres méthodes d'ordre 2 permettant d'assurer l'indépendance des sources, comme la méthode SOBI [BACEM97]. Ces méthodes sont fondées sur la diagonalisation jointe de matrices de covariance et sont utilisées notamment dans le cas de sources non stationnaires.

Pour terminer cette introduction à l'analyse en composantes indépendantes, disons que trois propriétés statistiques des sources peuvent être exploitées : la non-gaussiannité, la non-stationnarité ou la corrélation temporelle [Car01].

1.1.3 Cas sous-déterminé : moins de capteurs que de sources

Dans le cas où il y a moins de sources que de capteurs, on a affaire à un manque d'information. Une manière d'estimer la matrice de mélange est alors l'utilisation de l'algorithme EM dans lequel les sources sont les paramètres cachés (sur lesquels on met une loi *a priori* super-gaussienne, en général une loi de Laplace). On obtient une estimation de la matrice de mélange au maximum de vraisemblance par un algorithme itératif. Cet algorithme fournit aussi une estimation des sources, à chaque étape, sous la forme d'une espérance conditionnelle : $E[S|X, A^k]$.

Estimation par l'algorithme EM

On se place maintenant dans le cas de séparation de sources bruitées [BC99a], où il y a éventuellement moins de capteurs (m capteurs, $m > 1$) que de sources (n sources). L'équation de mélange s'écrit alors :

$$x(t) = As(t) + b(t), \tag{1.15}$$

où $b \sim \mathcal{N}(0, J)$ est un bruit, éventuellement coloré, gaussien, centré.

Notez que dans toute la suite, on supposera que les sources sont centrées, ce qui n'est pas restrictif. Il suffit pour cela d'ôter aux observations leur moyenne, étant donnée la relation linéaire qui existe entre les sources et les observations.

Dans ce cadre sous-déterminé, il est assez naturel d'utiliser un algorithme de type Expectation-Maximization [DLR77] pour l'estimation de la matrice de mélange ainsi que les sources, au maximum de vraisemblance.

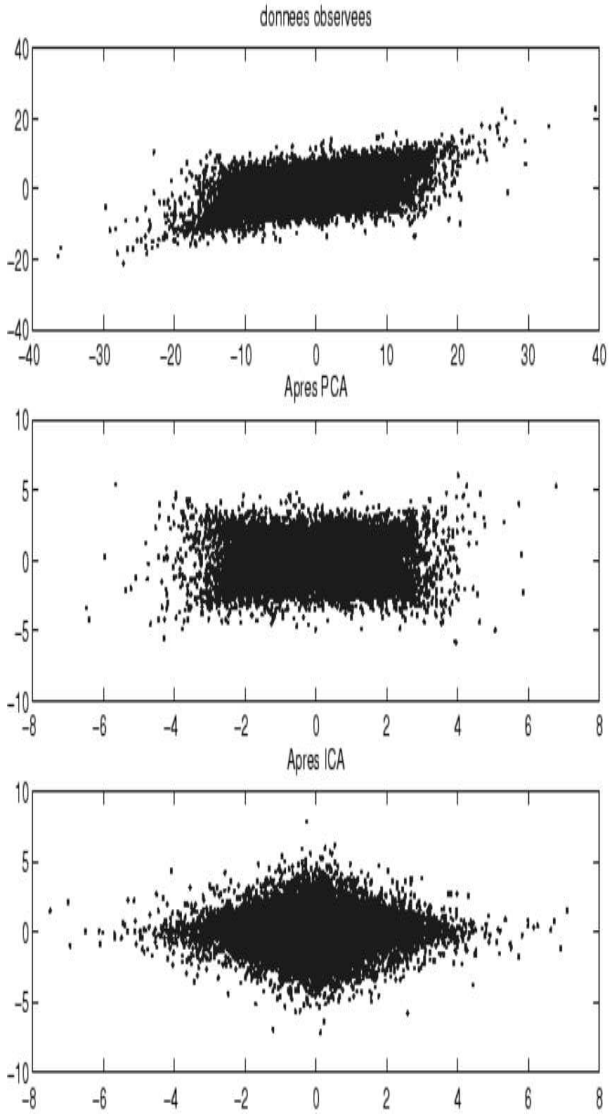

FIG. 1.1 – Exemple de séparation de sources laplaciennes

En effet, si l'on considère que la matrice de mélange et la covariance du bruit sont les paramètres à estimer, alors on a affaire à des observations incomplètes car les sources sont inconnues.

La vraisemblance des données observées connaissant les paramètres A et J est :

$$p(x|A,J) = \frac{1}{(2\pi)^{n/2}\sqrt{|\det J|}} \int_s r(s) \exp[-\frac{1}{2}(x-As)^T J^{-1}(x-As)]\, ds \tag{1.16}$$

où $r(s)$ est la densité *a priori* des sources. Dans le cas de sources indépendantes, cette densité est factorisable.

On a utilisé ici la loi des observations, sachant la matrice de mélange et la covariance du bruit. Cette loi est marginale par rapport aux sources, d'où l'intégrale dans l'expression (1.16). En effet, la densité des données complètes (x, s), s'écrit :

$$p(x, s|A, J) = \frac{1}{(2\pi)^{n/2}\sqrt{|\det J|}} r(s) \exp[-\frac{1}{2}(x - As)^T J^{-1}(x - As)] \,. \tag{1.17}$$

L'algorithme EM permet de trouver un maximum local pour la densité marginale, en itérant deux étapes :

- une étape E (Expectation) : on calcule l'espérance de fonctions (statistiques exhaustives) des données non observées.
- une étape M (Maximization) : on estime les nouvelles valeurs des paramètres comme si les statistiques sur les données complètes étaient observées.

En langage mathématique, cela s'écrit

- Etape E : calculer $Q(\theta, \theta^k) = E[\log p(x, s|\theta)|x, \theta^k]$,
- Etape M : maximiser $\theta^{k+1} = \arg\max_\theta Q(\theta, \theta^k)$.

Dans le cas de la séparation de sources bruitées, l'étape M est simple. Après calculs, on obtient :

$$A^{k+1} = R_{xs}^k {R_{ss}^k}^{-1} \tag{1.18}$$

$$J^{k+1} = R_{xx} - R_{xs}^k {R_{ss}^k}^{-1} {R_{xs}^k}^T, \tag{1.19}$$

où l'on a dénoté les moments empiriques

$$R_{xs}^k = \frac{1}{T}\sum_{t=1}^{T} x(t) E[s(t)^T|x, A^k, J^k] \tag{1.20}$$

$$R_{ss}^k = \frac{1}{T}\sum_{t=1}^{T} E[s(t)s(t)^T|x, A^k, J^k], \tag{1.21}$$

et $R_{xx} = \frac{1}{T}\sum_{t=1}^{T} x(t)x(t)^T$ la covariance empirique des observations.

Toute la difficulté réside dans le calcul de ces moments (Étape E). Une option serait de calculer les espérances par une méthode de Monte-Carlo. Cependant, cela devrait être fait à chaque étape, pour chaque index de temps t. En pratique, c'est beaucoup trop coûteux algorithmiquement.

Au contraire, une option minimale consiste à remplacer les espérances sur $s(t)$ par la valeur au maximum de vraisemblance $\hat{s}(t)$

$$\hat{s}(t)^k = \arg\max_s \left[\log r(s) - \frac{1}{2}(x(t) - A^k s)^T J^{k-1}(x(t) - A^k s)\right]. \tag{1.22}$$

Pour aller un peu plus loin, on peut utiliser, au premier ordre, une approximation de Laplace. On écrit pour cela la Hessienne de la densité complète au point $\hat{s}(t)$:

$$H(s) = -\frac{\partial^2}{\partial s^2} \log p(x, s|A, J) \tag{1.23}$$

$$= A^T J^{-1} A - \frac{\partial^2}{\partial s^2} \log r(s) \,. \tag{1.24}$$

On en déduit alors l'approximation gaussienne autour du maximum $\hat{s}$, suivante :

$$p(s|A, J, x) \approx \frac{1}{(2\pi)^{m/2}\sqrt{|\det H(\hat{s})|}} \exp[-\frac{1}{2}(s-\hat{s})^T H(\hat{s})(s-\hat{s})] \,. \tag{1.25}$$

D'où finalement le résultat :

$$E[s(t)^T|x, A^k, J^k] = \hat{s} \tag{1.26}$$

$$E[s(t)s(t)^T|x, A^k, J^k] = \hat{s}\hat{s}^T + H(\hat{s})^{-1} \,. \tag{1.27}$$

Une autre option pour estimer les moments empiriques consiste à modéliser la densité *a priori* sur les sources $r(s)$ par un modèle de mélange de gaussienne, modèle paramétrique. En effet, dans le cas gaussien, tel qu'on l'a vu précédemment, les calculs sont simples et s'étendent naturellement aux mélanges de gaussiennes dans un formalisme de données incomplètes [BMC97].

Digression sur les représentations parcimonieuses

Nous avons vu que l'algorithme EM nécessite en pratique l'estimation des sources au maximum *a posteriori*. Nous allons nous attarder un peu ici sur cet aspect là de l'estimation, afin de faire le lien avec les représentations parcimonieuses.

Nous repartons donc de l'équation (1.22) d'estimation des sources. Dans le cas d'une distribution *a priori* $r(s)$ à longue queue et dans le cas d'un nombre plus grand de sources que de capteurs, i.e. $m > n$, on peut obtenir ce que l'on appelle une représentation parcimonieuse de $x(t) \in \mathcal{R}^n$ à travers le dictionnaire $A \in \mathcal{R}^{n\times m}$ ($s(t) \in \mathcal{R}^m$). Les colonnes de la matrice A sont alors appelés *atomes*.

En particulier, dans le cas des gaussiennes généralisées, c'est-à-dire lorsque la densité est de la forme $-\log r(s) = \frac{\gamma}{2} \sum_{i=1}^{m} |s_i(t)|^\alpha$, avec $\alpha \leq 1$ [KDRE99], l'équation (1.22) fournit des solutions parcimonieuses, c'est-à-dire ayant au plus n coefficients de sources non nuls à chaque temps t et ayant de bonnes propriétés de compacité de la représentation. Pour expliquer cette notion de compacité, on peut dire par exemple qu'il est plus efficace de représenter une sinusoïde sur une base d'exponentielles complexes (où il y aura deux coefficients non nuls) qu'à partir de diracs temporels (base canonique) où chaque composante sera nécessaire.

Le problème d'estimation des sources au maximum de vraisemblance dans le cas de gaussiennes généralisées devient

$$\hat{s}(t) = \arg\min_{s} \|x(t) - As\|_2^2 + \gamma \|s\|_\alpha^\alpha, \tag{1.28}$$

dans le cas d'une matrice de covariance du bruit J diagonale.

Nous utiliserons ces représentations parcimonieuses dans la suite de cette thèse, dans le cadre de représentations non-négatives du spectre : c'est pourquoi nous introduisons cette notion dès à présent.

1.2 Séparation de sources sonores avec un seul capteur

Ce domaine commence à être actif depuis les années 2000. Une proportion importante de chercheurs à avoir abordé ce problème provient du domaine du traitement de la parole, notamment dans le cadre de la reconnaissance de la parole en milieu bruité. Par ailleurs, depuis que la séparation de sources est devenue un domaine à part entière en traitement du signal, en particulier grâce aux succès de la communauté ACI ainsi qu'aux questions soulevées par l'approche CASA, quelques chercheurs se sont attaqués au problème spécifique de la séparation avec un seul capteur. Leurs approches sont différentes : adaptation de méthode de type ACI, utilisation de modèles prédictifs ou encore modélisation par mélange de lois sur des données spectrales.

Nous commencerons par présenter, parmi ce foisonnement d'idées, les propositions faites par S. Roweis, qui ont de nombreux points communs avec un des algorithmes qui seront développés dans la suite de ce travail.

1.2.1 Les propositions de Sam Roweis

Le procédé le plus proche de ce que nous proposerons dans le cadre de modèles de mélange de lois est décrit dans l'article *One microphone source separation* [Row00]. L'auteur y décrit un système permettant la séparation de deux sources à l'aide d'un seul microphone. Il utilise des bancs de filtres pour extraire des caractéristiques spectrales $x_k(t)$ associées à la bande de fréquence k et au temps t (représentation temps-fréquence et réduction de la dimension). Il utilise également des Modèles de Markov Cachés pour chacune des sources et chacune des bandes fréquentielles k. C'est l'utilisation de Modèles de Markov cachés pour chaque source, dans le domaine temps-fréquence (réduit) et le calcul d'un filtre binaire (c'est-à-dire prenant comme seules valeurs possibles 0 ou 1) résultant de la superposition des sources qui caractérise son approche.

A chaque état ou composante du modèle de Markov caché est associée une moyenne log-spectrale $m_j^i(k)$ où i est l'indice de la source ($i = 1, 2$ pour deux sources), j est l'indice de la composante ou état actif et k est l'indice du filtre fréquentiel utilisé.

Roweis déduit une formule de calcul des caractéristiques log-spectrales du processus mélangé $x = s_1 + s_2$, somme des deux sources. Cette formule associe la moyenne liée au couple d'états actifs (j_1, j_2) pour chacun des deux Modèles de Markov Cachés (HMM ou Hidden Markov Models en Anglais) et à la bande de fréquence k, à chacune des moyennes log-spectrales des deux HMM (pour les états respectifs j_1, j_2 et la même bande de fréquence k) :

$$m_{j_1,j_2}(k) = \max\{m_{j_1}^1(k),\, m_{j_2}^2(k)\}\,. \tag{1.29}$$

La connaissance des caractéristiques spectrales des HMM résultant de la somme des deux sources (HMM dits factoriels car résultants de la factorisation des HMM de chacune des deux sources, pour chacune des bandes de fréquences) permet de calculer le couple d'états $j_1(t,k), j_2(t,k)$ le plus probable pour chaque indice de temps t et chaque bande de fréquence k (décodage de Viterbi).

Roweis utilise ensuite des masques binaires pour reconstruire les sources :

$$s_1(t) = a_1(t)x_1(t) + a_2(t)x_2(t) + \ldots + a_n(t)x_n(t), \tag{1.30}$$

a_k représente le masque pour la bande de fréquence k et $x_k(t)$ représente la caractéristique spectrale du signal composite (mélange capté par le microphone) en sortie du banc de filtre numéro k.

Les masques, binaires, sont calculés de la manière suivante :

$$a_k(t) = \begin{cases} 1 & \text{si } m_{j_1(t,k)}(k) > m_{j_2(t,k)}(k) \\ 0 & \text{sinon} \end{cases}\,. \tag{1.31}$$

où $j_1(t,k)$ et $j_2(t,k)$ sont les états décodés par le HMM factoriel associé à la bande de fréquence k, au temps t.

Nous verrons plus tard une justification et quelques précisions sur les propositions de Roweis à partir d'une ré-interprétation du filtrage de Wiener avec des modèles *a priori* sur les sources sous formes de mélanges de gaussiennes ou de HMM à densité conditionnelle gaussienne.

1.2.2 Le débruitage des signaux de parole

Dans le cadre de la communauté du traitement de la parole, le problème posé consiste plutôt dans le débruitage des signaux en vue de traitement ultérieur (reconnaissance, par

exemple). Comme la modélisation par mélange de lois de type HMM est classique en traitement de la parole, ce type d'approche a été privilégié. Ces travaux sont parmi les plus "anciens", par rapport à l'intérêt suscité par notre problématique. Citons donc les travaux de Y. Ephraim [Eph92], Varga et Moore [VM90].

Les travaux de Yariv Ephraim sont assez proches d'un point de vue théorique de ce que nous allons développer dans le cadre de l'estimateur de l'espérance conditionnelle avec des Modèles de Mélange de Gaussiennes (MMG), à ceci près qu'Ephraim utilise des modèles type "switching AR", qui conduisent à des calculs difficiles et sont moins concluants d'un point de vue perceptuel, par rapport à un travail effectué directement sur le spectrogramme. Nous entendons par Switching AR, un modèle à état discret, auquel correspond un modèle de prédiction linéaire du signal à chaque état. Le spectre (DSP) est donc estimé de manière paramétrique.

Les travaux de Varga et Moore sont proches de nos orientations, puisqu'ils font appel à une modélisation avec HMM sur le logarithme du spectre de puissance. La différence se situe dans le fait qu'ils ne restaurent pas le signal, mais se limitent à un décodage conjoint des deux HMM (l'un pour la parole, l'autre pour le bruit) dans l'objectif de reconnaissance de la parole. Ces travaux ont aussi en commun avec notre approche l'utilisation d'un apprentissage préalable sur des sources séparées.

1.2.3 L'approche en sous-espaces indépendants dans le domaine spectral

Cette approche (Independent Subspace Analysis ou ISA) se fonde sur une décomposition du spectrogramme du signal observé en sous-espaces indépendants. Le spectrogramme est décomposé en une somme d'atomes temps-fréquence séparables :

$$\mathcal{S}x(t,f) = u_1(f) \cdot a_1(t) + \ldots + u_n(f) \cdot a_n(t).$$

Cela revient en fait à décomposer le spectre en atomes fréquentiels de base $u_i(f)$ modulés dans le temps par une fonction d'activation $a_i(t)$. Casey et Westner [CW00] utilisent une décomposition en valeurs singulières pour obtenir cette factorisation : dans ce cas, les atomes fréquentiels sont orthogonaux. Les auteurs regroupent ensuite les divers atomes pour former les sources et la reconstruction se fait par simple inversion du spectrogramme ou par projection. Cette approche (ISA) ressemble un peu à ce que nous proposons par la suite avec des décompositions non-négatives du spectre de puissance. Cependant, nous ne travaillerons pas directement sur le spectrogramme mais sur son module au carré, afin d'éviter des difficultés liées à la phase des signaux. La reconstruction sera donc naturellement différente. D'autre part, dans l'approche ISA, on n'utilise pas ou peu d'information *a priori* sur les sources.

1.2.4 Les approches de type ICA

Dans un article prévu pour la conférence ICASSP 2003 [JLO03], les auteurs décrivent une nouvelle méthode de séparation de sources avec un seul capteur, dans la lignée de l'Analyse en Composantes Indépendantes (ACI).

Le premier point est qu'ils décomposent le mélange sur une base adaptée, apprise par ACI sur un corpus représentatif des différentes sources en présence. L'originalité de ce travail est que chaque composante, sur cette base et pour chaque source, est modélisée par une gaussienne généralisée ($\log p(s_k) \propto |s_k|^{\alpha_k}$) et c'est précisément le coefficient α_k qui est estimé dans une phase d'apprentissage, pour chaque source et chaque vecteur de base.
Il en résulte une décomposition du signal composite (mélange) au moment de la séparation en deux signaux, au maximum de vraisemblance. Cela s'apparente à un filtre de Wiener généralisé et la résolution du problème nécessite une étape de descente du gradient. De manière grossière, on peut dire que cette méthode revient à sélectionner dans un dictionnaire appris sur un corpus audio, des composantes propres à l'une des sources et pas aux autres. Cela s'apparente à la méthode de réunion des dictionnaires adaptés que nous avions proposée au GRETSI en 2001 [BGB01].

En ce qui concerne les performances de cette méthode, elles sont difficiles à comparer à celles que nous présentons dans ce manuscrit, car les critères d'évaluation diffèrent. Cela dit, dans les deux approches, il y a une généralisation du filtrage de Wiener et il serait intéressant d'essayer de comprendre les connexions entre ces approches liées à l'ACI et les approches par mélange de lois.

1.2.5 Les modèles prédictifs

Une alternative à la paramétrisation liée au spectrogramme provient de l'utilisation de modèles prédictifs plus ou moins élaborés. Le cadre naturel pour la restauration ou la séparation de sources n'est plus le filtrage de Wiener mais le filtrage de Kalman. Dans l'article [WN97], les auteurs utilisent des modèles prédictifs à base de réseaux de neurones, donc non-linéaires, et utilisent deux filtres de Kalman en parallèle pour chacune des sources. Les auteurs semblent penser que leur méthode pourrait permettre une séparation de sources avec un seul capteur, sans apprentissage, bien qu'ils restent prudents sur cette éventualité.

On peut dire que la modélisation proposée par Wan et Nelson est une alternative à la modélisation par mélange de lois sur le spectrogramme et que l'on retrouve naturellement ces deux possibilités dans la théorie de l'estimation spectrale pour des signaux non-stationnaires [Cas03], c'est-à-dire les méthodes paramétriques (modèles auto-régressifs, par exemple) et les

méthodes non paramétriques (spectrogramme).

1.2.6 Conclusion

Pour conclure, on peut dire que chaque méthode proposée apporte sa contribution à ce nouveau problème et que les orientations choisies dépendent souvent de la communauté à laquelle appartiennent les auteurs (parole ou ACI par exemple). Il est encore difficile à l'heure actuelle de se faire une idée générale sur la question, aussi bien que de comparer suivant des critères identiques les différents algorithmes proposés. Cependant, il semble que l'on puisse classifier les différentes méthodes suivant deux critères : la méthode d'estimation spectrale utilisée (paramétrique/ non-paramétrique) et l'utilisation ou l'absence de modèle à état discret (Modèle de Mélange de Gaussiennes, Modèle de Markov Cachés, etc). En revanche, la distinction filtre de Wiener/ Kalman nous paraît plutôt provenir de la façon de poser le problème et donc des deux distinctions précédentes, que d'une véritable différence dans les traitements.

Chapitre 2

Hypothèse de stationnarité locale

2.1 Stationnarité locale

Dans tout notre travail, nous poserons une hypothèse fondamentale sur les signaux audio que nous étudions : la stationnarité à court terme. Nous tentons donc ici de donner une définition de la stationnarité locale, qui justifiera l'utilisation de la transformée de Fourier à court terme, pour l'estimation de densités spectrales de puissance. Cette hypothèse est à la base de l'utilisation de matrices de covariance diagonales dans le domaine de Fourier dans nos modèles de sources et nous allons essayer de justifier cette propriété dans ce chapitre. On pourra consulter la thèse [Coa98] pour plus d'information sur les notions temps-fréquence.

2.1.1 Spectre variant dans le temps

Nous allons définir la notion de spectre variant dans le temps (en anglais : time-varying spectrum), pour un processus aléatoire $X(t)$ centré de covariance :

$$R(t,s) = \mathrm{E}\{X(t)X^{\star}(s)\}\,. \tag{2.1}$$

L'opérateur de covariance est défini pour toute fonction de carré sommable f, par :

$$Tf(t) = \int R(t,s)f(s)ds\,. \tag{2.2}$$

On peut ré-écrire la covariance en fonction de $t-s$ et $\frac{t+s}{2}$ (changement de variable), d'où :

$$R(t,s) = C_0(\frac{t+s}{2}, t-s)\,. \tag{2.3}$$

Pour un processus stationnaire, la covariance ne dépend que de $t-s$:

$$C_0(\frac{t+s}{2}, t-s) = C_0(t-s), \tag{2.4}$$

et l'opérateur de covariance est une convolution

$$Tf(t) = \int C_0(t-s)f(s)ds = C_0 \star f(t)\,. \tag{2.5}$$

Avec le changement de variable (2.3), l'opérateur de covariance

$$Tf(t) = \int C_0(\frac{t+s}{2}, t-s)f(s)ds \tag{2.6}$$

peut être interprété comme une convolution variant dans le temps.
Si $C_0(u,v)$ varie lentement en fonction de u, on définit le spectre variant dans le temps (time varying spectrum, en anglais) par la Transformée de Fourier de $C_0(u,v)$ en fonction de v :

$$\Lambda_0(u,\omega) = \int C_0(u,v)e^{-i\omega v}dv\,. \tag{2.7}$$

$\Lambda_0(u,\omega)$ est encore appelé symbole de Weyl. Bien que l'on parle de *spectre* variant dans le temps, $\Lambda_0(u,\omega)$ peut être négatif pour certaines valeurs particulières de u et ω. Il s'agit en fait de l'espérance de la distribution de Wigner-Ville du processus $X(t)$:

$$\Lambda_0(u,\omega) = \mathrm{E}\{WX(t)\}, \tag{2.8}$$

où la distribution de Wigner-Ville est définie par :

$$Wf(u,\omega) = \int f(u+\frac{v}{2})f^\star(u-\frac{v}{2})e^{-i\omega v}dv \tag{2.9}$$

Cela généralise donc la notion de densité spectrale de puissance dans le cas stationnaire, puisque celle-ci est l'espérance du spectre de puissance : $\mathrm{E}\{|X(f)|^2\}$.

2.1.2 Définition de la stationnarité locale

Pour un processus localement stationnaire, à chaque temps t_0, il existe un intervalle de longueur $l(t_0)$, où le processus est approximativement stationnaire. Autrement dit, pour tous $t, s \in [t_0 - \frac{l(t_0)}{2}, t_0 + \frac{l(t_0)}{2}]$, la covariance ne dépend (approximativement) que de $t-s$:

$$\mathrm{E}\{X(t)X^\star(s)\} \approx C(t_0, t-s), \text{ si } |t-s| < \frac{l(t_0)}{2}\,. \tag{2.10}$$

Par ailleurs, on définit la distance maximale d de corrélation entre deux points (decorrelation length, en anglais) :

$$\mathrm{E}\{X(t)X^\star(s)\} \approx 0, \text{ si } |t-s| > d\,. \tag{2.11}$$

Pour un processus localement stationnaire, on considère que l'on a $d < \frac{l(t_0)}{2}$, pour tout t_0.

Après cette définition, donnons la propriété principale d'un processus localement stationnaire. Soit, pour cela, une fenêtre glissante, régulière, $g_{t_0}(t)$, dont le support est précisément $[t_0 - \frac{l(t_0)}{2}, t_0 + \frac{l(t_0)}{2}]$ (centrée en t_0) et soit l'atome temps-fréquence $\phi_{t_0,\xi}(t) = g_{t_0}(t) \cdot e^{i\xi t}$. Alors, si $X(t)$ est un processus localement stationnaire, on a approximativement :

$$T\phi_{t_0,\xi}(t) \approx \Lambda_0(t_0, \xi)\phi_{t_0,\xi}(t), \tag{2.12}$$

c'est-à-dire que l'atome $\phi_{t_0,\xi}(t)$, dont l'énergie est centrée autour du point (t_0, ξ) dans le plan temps-fréquence, est approximativement vecteur propre de l'opérateur de covariance du processus $X(t)$, avec la valeur propre $\Lambda_0(t_0, \xi)$.

Cette valeur propre est assimilable à une densité spectrale de puissance locale. Dans le cas où cette même valeur propre intervient à différents instant t_i, pour chaque fréquence ξ, on peut estimer cette densité spectrale de puissance en moyennant le spectre du puissance "local", à ces différents instants, ce spectre étant estimé par la transformée de Fourier à court-terme de l'observation $x(t)$.

Autrement dit, s'il existe $t_1, \ldots, t_n$ tels que

$$\forall i,\ \forall \xi,\ \Lambda_0(t_i, \xi) = D(\xi), \tag{2.13}$$

alors $D(\xi)$ (la DSP locale) peut être estimée par :

$$D(\xi) \approx \frac{\sum_i |\mathcal{S}x(t_i, f)|^2}{n}, \tag{2.14}$$

où $\mathcal{S}$ est l'opérateur de la Transformée de Fourier à Court Terme (TFCT, ou Short-Time Fourier Transform, an Anglais (STFT)).

Notez que théoriquement la longueur de l'intervalle de stationnarité du processus $l(t_0)$ dépend du temps t_0 (voir [MPZ98]). Cependant, nous utiliserons des fenêtres de taille constante.

2.2 Utilisation de la Transformée de Fourier à Court Terme

La Transformée de Fourier à Court Terme est redondante et elle permet de passer du domaine temporel au domaine temps-fréquence suivant une relation linéaire :

$$\mathcal{S}x(t, f) = \sum_{\tau=-N/2}^{N/2-1} \exp[2\pi i f \frac{\tau}{N}] \cdot \omega(\tau) \cdot x(t + \tau), \tag{2.15}$$

où $\omega(\tau)$ est la fenêtre d'analyse et N est la taille de cette fenêtre (typiquement 20 millisecondes, soit quelques centaines d'échantillons pour des signaux échantillonnés à 11kHz). Nous avons

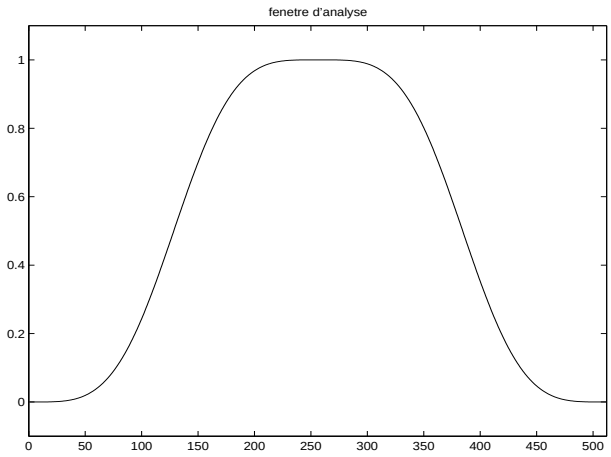

FIG. 2.1 – Fenêtre d'analyse $\omega(\tau)$ sur 512 points

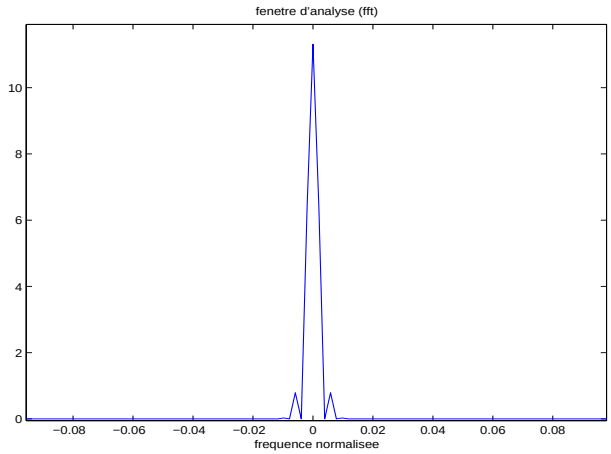

FIG. 2.2 – Module de la transformée de Fourier de la fenêtre d'analyse $\omega(\tau)$

utilisé comme fenêtre d'analyse : $\omega(\tau) = \frac{1+\sin\left(\frac{\pi}{2}\cos\left(\pi\frac{2\tau}{N}\right)\right)}{2}$. Ce type de fenêtre est classiquement utilisé dans les analyses en ondelettes de Malvar (Local cosine bases, en Anglais) [WW93].

Notez que l'on a échantillonné la TFCT avec un pas de $N/2$ points (décalage de la fenêtre d'analyse), ce qui fait que la reconstruction est directe et s'obtient par addition des différentes portions de signal fenêtrées.

Dans tout ce qui suit, nous avons utilisé la TFCT sur 512 points à 11kHz, soit des fenêtres de 47 millisecondes environ.

Problème de la reconstruction des sources

Comme on utilise une transformée redondante et que l'on va opter pour une méthode de filtrage non-linéaire pour la séparation, un problème se pose à la reconstruction, pour les sources estimées. En effet la TFCT, notée $\mathcal{S}$, étant redondante, il n'est pas obligatoire qu'après filtrage on se trouve dans $\mathcal{I}\mathrm{m}\mathcal{S}$ (l'image de l'espace signal par la TFCT, qui est un espace vectoriel beaucoup plus grand que l'espace de départ, l'espace signal, et cela du fait du recouvrement des trames) (voir figure 2.3). En conséquence, il faut projeter le signal temps-fréquence filtré sur $\mathcal{I}\mathrm{m}\mathcal{S}$, ce qui revient à chercher la source dans l'espace signal qui a la TFCT la plus proche de celle estimée (filtrée), au sens de la norme L_2.

2.3 Conclusion

L'hypothèse de stationnarité locale des sources nous conduit à utiliser une transformée temps-fréquence linéaire telle que la transformée de Fourier à court terme et nous pouvons alors considérer que les matrices de covariances de nos modèles, par exemple pour des modèles à état, sont diagonales dans le domaine fréquentiel. Ces matrices approximent des Densités Spectrales de Puissance locales.

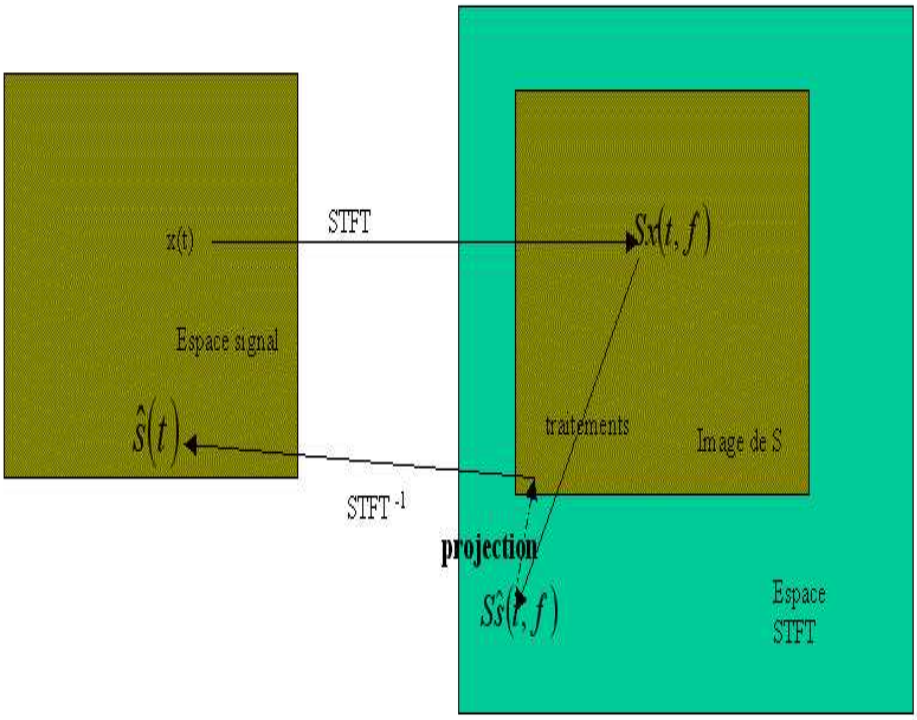

FIG. 2.3 – Projection du signal dans $\mathcal{I}\mathrm{m}\mathcal{S}$, après traitement

Chapitre 3

Problématique de l'évaluation

La définition de critères d'évaluation est une étape importante pour comparer différents algorithmes de séparation de sources, ce que nous ferons dans la partie IV.

Les critères généralement utilisés dans le cas déterminé ou sur-déterminé (plus de capteurs que de sources) concernent la matrice de mélange. Ces critères ne sont donc pas transposables à notre étude.

D'autre part, il paraît logique d'utiliser des critères du même type que ceux utilisés en débruitage, c'est-à-dire le plus souvent un rapport signal à bruit exprimé en décibels. Cependant, nous trouvons dans une situation légèrement différente du débruitage, car nous avons la possibilité de prendre plusieurs signaux de référence, à savoir les différentes sources de départ, alors que dans le cas du débruitage le seul signal original importe.

Cette problématique de l'évaluation de la séparation de sources dans le cas sous-déterminé a fait l'objet de différents travaux [STS99] dont ceux d'un groupe de travail du GDR ISIS [GBVF03], auquel l'auteur de ces lignes a participé. Dans le cas de deux sources et un seul capteur, nous avons proposé deux critères pour chacune des sources estimées. Ces critères sont essentiellement des rapports de norme ℓ_2 de signaux (distance euclidienne).

Le premier critère est fondé sur l'interférence entre la source originale recherchée et l'autre source originale que l'on voulait atténuer, voire éliminer. Nous avons appelé ce critère SIR pour Source to Interference Ratio.

Le second critère concerne le bruit additif, c'est à dire les distortions générées par l'algorithme de séparation, qui est nécessairement non-linéaire. Ce second critère est le SAR ou Source to Artefact Ratio. Nous donnerons une définition précise de ces critères d'évaluation dans la partie IV qui est consacrée à l'étude expérimentale.

Cependant, il demeure plusieurs points noirs dans cette problématique de l'évaluation :

– Le premier concerne l'utilisation de critères numériques. L'utilisation d'une distance eu-

clidienne est probablement ce qu'il y a de plus simple d'un point de vue mathématique, mais perceptivement, cela peut être décevant, car des signaux différents risquent d'obtenir des scores voisins et, inversement, des signaux perceptivement identiques peuvent obtenir des scores différents. L'utilisation de critères psycho-acoustiques pour l'évaluation (mais aussi dans l'estimation des sources, au niveau de la fonction de coût à optimiser) reste une perspective non-explorée de cette thèse.

– La seconde critique concerne aussi l'utilisation de la norme ℓ_2 dans les critères, mais d'un point de vue mathématique. L'utilisation de cette norme induit un moyennage et est donc peu robuste aux données aberrantes. On prefèrerait un critère plus "local", car la séparation peut être bonne en moyenne, mais catastrophique dans une zone temporelle réduite. En fait, les critères d'évaluation dépendent fortement de l'application visée et de ses exigences ; ils devraient être donc définis dans cette optique.

Deuxième partie

Approche probabiliste bayésienne

Cette partie est l'occasion d'introduire la modélisation probabiliste des sources sonores et de donner un cadre théorique cohérent pour l'estimation de ces sources à partir d'un mélange : le cadre bayésien. L'approche probabiliste nous permet de tenir compte des variabilités au sein d'une source sonore, qui est considérée comme une réalisation d'un processus aléatoire complexe. Le cadre bayésien permet d'introduire naturellement la notion de connaissances *a priori* sur les sources sonores, qui correspond à une phase d'apprentissage afin d'extraire des caractéristiques de base de chacune des sources. Dans la phase de séparation, on utilise la notion de loi *a posteriori* pour estimer les sources en fonction du mélange observé et des connaissances provenant de la phase d'apprentissage.

Le premier chapitre est un rappel théorique concernant la théorie bayésienne, notamment dans le cadre de l'estimation ponctuelle.

Le second chapitre est une application directe du paradigme précédent à la séparation de sources, dans le cas de modèles simplifiés pour les sources. A partir de la théorie bayésienne, on redémontre ici les formules du filtrage de Wiener, qui serviront de référence dans le cadre de l'évaluation des méthodes dans la partie IV.

Le troisième chapitre étend le filtre de Wiener à des modèles plus réalistes de sources sonores. La modélisation est fondée sur des modèles à états cachés tels que les modèles de mélange de gaussiennes ou les modèles de Markov cachés à densités conditionnelles gaussiennes.

Le quatrième chapitre traite de la modélisation du logarithme du spectre de puissance des signaux, avec des modèles MMG.

Chapitre 4

Formalisme bayésien

Nous rappelons dans ce chapitre quelques notions sur la théorie bayésienne, car elle constitue la base des développements que nous proposons sur le filtrage de Wiener.

4.1 Modèle bayésien

Un modèle bayésien se caractérise par la donnée d'un modèle paramétrique et d'une loi *a priori* sur les paramètres.
Un *modèle paramétrique* consiste en l'observation d'une variable aléatoire x, distribuée selon $f(x|\theta)$, où seul le paramètre θ est inconnu et appartient à un espace vectoriel de dimension finie.
Un *modèle statistique bayésien* est alors constitué d'un modèle statistique paramétrique, $f(x|\theta)$, appelé fonction de vraisemblance, et d'une distribution *a priori* sur les paramètres $\pi(\theta)$.

Remarque : en pratique, nous ne connaissons pas la vraie loi selon laquelle la variable aléatoire x est générée, même à un jeu de paramètres près, si tant est que cette loi existe. En revanche, la loi paramétrique $f(x|\theta)$ tente d'approcher cette loi idéale et la distribution $\pi(\theta)$ représente la connaissance *a priori* que l'on a sur le paramètre θ.
Cette définition d'un modèle bayésien conduit à la notion de loi *a posteriori*

$$p(\theta|x) \propto f(x|\theta)\pi(\theta)\,.$$

C'est à partir de cette loi *a posteriori* que l'inférence de θ va être possible et en ce sens, cette notion de loi *a posteriori* joue un rôle central dans la théorie bayésienne. Donnons la définition d'une autre quantité importante : la loi marginale

$$m(x) = \int_\theta f(x|\theta)\pi(\theta)d\theta\,.$$

4.2 Fonction de coût et inférence bayésienne

L'estimation du paramètre θ à partir d'une observation x de la variable aléatoire, nécessite la définition d'une fonction de coût $C(\delta, \theta)$. Cette fonction $C(\delta, \theta)$ représente le coût du remplacement de la valeur vraie du paramètre θ par son estimée δ.
L'estimation de θ revient à minimiser le coût moyen, sur l'ensemble des valeurs possibles du paramètre θ :

$$\delta_{opt} = \arg\min_{\delta} \int_{\theta} C(\delta, \theta) f(x|\theta) \pi(\theta) d\theta \; .$$

Dans le cas où la fonction de coût est la distance quadratique entre la valeur du paramètre θ et son estimée δ : $C(\delta, \theta) = \|\delta - \theta\|_2^2$, alors l'estimateur bayésien δ_{opt} est bien connu : c'est l'espérance conditionnelle $E(\theta|x)$:

$$E(\theta|x) = \frac{\int_{\theta} \theta f(x|\theta) \pi(\theta) d\theta}{\int_{\theta} f(x|\theta) \pi(\theta) d\theta} \; .$$

Il existe une autre fonction de coût standard : $C(\delta, \theta) = 1 - Dirac(\delta - \theta)$. Dans ce cas, l'estimateur bayésien est l'estimateur du maximum *a posteriori* (MAP) :

$$\delta_{opt} = \arg\max_{\theta} f(x|\theta) \pi(\theta) \; .$$

Pour plus de détails sur la théorie bayésienne et notamment sur les estimateurs ponctuels bayésiens, on pourra se reporter au livre [Rob01].

Chapitre 5

Filtrage de Wiener

Supposons que l'on considère deux processus gaussiens centrés $s_1(t)$ et $s_2(t)$ (les sources) et que l'on observe la somme bruitée de ces deux processus $x(t) = s_1(t) + s_2(t) + b(t)$ où b est un bruit blanc gaussien de variance σ_b^2.
Le problème consiste à estimer les deux processus en fonction de l'observation $x(t)$. Il s'agit bien d'un problème de séparation de sources avec un capteur unique.

Si l'on suppose que les processus sources sont stationnaires, alors leur matrice de covariance est diagonale dans la base de Fourier. On notera $\mathcal{F}$ l'opérateur associé à la Transformée de Fourier Discrète et $\sigma_1^2(f)$, $\sigma_2^2(f)$ les diagonales des matrices de covariance de ces deux processus dans cette base (f désigne ici la fréquence). $\sigma_i^2(f)$, avec $i = 1, 2$ sont appelées Densités Spectrales de Puissance. Ce sont les données *a priori* du problème.

5.1 Modèle bayésien

On peut, à partir des données de ce problème, construire un modèle bayésien, qui nous permet d'estimer les sources dans le cadre de cette théorie. En effet, le modèle de bruit nous permet d'introduire la vraisemblance des données observées en fonction des sources, qui sont les paramètres à estimer. De même, les covariances de sources fournissent un modèle *a priori* sur ces sources. Le modèle porte sur la transformée de Fourier (complexe) des signaux. Cette transformée est linaire et conserve donc la relation additive entre mélange et sources :

$$\mathcal{F}x(f) = \mathcal{F}s_1(f) + \mathcal{F}s_2(f) + \mathcal{F}b(f), \tag{5.1}$$

et de plus,

$$\mathcal{F}b(f) \sim \mathcal{N}(0, \sigma_b^2), \tag{5.2}$$

$$\mathcal{F}s_1(f) \sim \mathcal{N}(0, \sigma_1^2(f)), \tag{5.3}$$

$$\mathcal{F}s_2(f) \sim \mathcal{N}(0, \sigma_2^2(f)). \tag{5.4}$$

La vraisemblance des données observées est donnée par la distribution du bruit et les processus s_1 et s_2 (ou plutôt leur transformée de Fourier Discrète $\mathcal{F}s_1(f)$ et $\mathcal{F}s_2(f)$) sont les paramètres :

$$p(x|s_1, s_2) \propto \exp\left[\sum_f \frac{|\mathcal{F}x(f) - \mathcal{F}s_1(f) - \mathcal{F}s_2(f)|^2}{2\sigma_b^2}\right] . \tag{5.5}$$

Les lois *a priori* résultent du modèle gaussien sur les sources :

$$p(s_1) \propto \exp\left[\sum_f \frac{|\mathcal{F}s_1(f)|^2}{2\sigma_1^2(f)}\right] , \tag{5.6}$$

$$p(s_2) \propto \exp\left[\sum_f \frac{|\mathcal{F}s_2(f)|^2}{2\sigma_2^2(f)}\right] . \tag{5.7}$$

Si on suppose de plus que les sources sont indépendantes, on obtient $p(s_1, s_2) = p(s_1) \cdot p(s_2)$. Par la loi de Bayes, on obtient la loi *a posteriori* :

$$p(s_1, s_2|x) \propto p(x|s_1, s_2) \cdot p(s_1, s_2) . \tag{5.8}$$

D'où finalement :

$$p(s_1, s_2|x) \propto \exp\left[\sum_f \frac{|\mathcal{F}x(f) - \mathcal{F}s_1(f) - \mathcal{F}s_2(f)|^2}{2\sigma_b^2} + \frac{|\mathcal{F}s_1(f)|^2}{2\sigma_1^2(f)} + \frac{|\mathcal{F}s_2(f)|^2}{2\sigma_2^2(f)}\right] . \tag{5.9}$$

5.2 Estimation des sources

Pour obtenir par exemple l'estimateur du maximum *a posteriori*, on annule la dérivée du logarithme de l'expression (5.9), par rapport aux paramètres à estimer. Après calcul, on obtient :

$$\widehat{\mathcal{F}s_1(f)} = \frac{\sigma_1^2(f)}{\sigma_1^2(f) + \sigma_2^2(f) + \sigma_b^2}\mathcal{F}x(f), \tag{5.10}$$

$$\widehat{\mathcal{F}s_2(f)} = \frac{\sigma_2^2(f)}{\sigma_1^2(f) + \sigma_2^2(f) + \sigma_b^2}\mathcal{F}x(f) . \tag{5.11}$$

L'inconvénient de ce filtrage est qu'il est purement fréquentiel alors que dans le cadre de sources sonores, on aimerait bien avoir un filtrage qui varie dans le temps à l'intérieur du domaine temps-fréquence. Le filtrage proposé ici est le filtrage de Wiener [Wie49] (avec $\sigma_b \to 0$).

Notons que les covariances diagonales $\sigma_i^2(f)$ peuvent être assimilées à des DSP qui décrivent le contenu fréquentiel des signaux. On voit bien que dans un morceau de musique, il y a plusieurs DSP correspondant à différents notes ou timbres, différentes hauteurs et amplitudes. On sent donc les limitations du filtre de Wiener où chaque source est décrite par une seule DSP. On va lever cette limitation en utilisant des modèles de mélanges de gaussiennes.

Chapitre 6

Extension aux Modèles de Mélange de Gaussiennes

Nous allons prolonger le filtre de Wiener, qui est optimal dans le cas de densités *a priori* gaussiennes, à des modèles de mélanges de gaussiennes. Cela nous permettra de lever l'hypothèse de stationnarité inhérente au filtrage de Wiener fréquentiel.

6.1 Introduction

Nous observons la superposition de deux sources $x(t) = s_1(t) + s_2(t)$ et nous avons un modèle *a priori* pour chacune des sources s_1 et s_2. Nous allons considérer la transformée de Fourier à court terme des signaux audio, afin de pouvoir exprimer la stationnarité locale de ces signaux. L'équation de mélange devient :

$$\mathcal{S}x(t,f) = \mathcal{S}s_1(t,f) + \mathcal{S}s_2(t,f) + \mathcal{S}b(t,f), \tag{6.1}$$

où $\mathcal{S}$ correspond à l'opérateur de transformée de Fourier à court terme.

Nous considérons que chaque trame de signal $\mathcal{S}s(t,f)$, à t fixé, est la réalisation d'un processus aléatoire et nous considérons que les trames d'un même signal sont des réalisations indépendantes, ce qui est une approximation en particulier du fait du recouvrement des trames.

Chaque trame de bruit $\mathcal{S}b(t,f)$, à t fixé, est considérée comme la réalisation d'un bruit blanc gaussien de variance σ_b^2.

$$\mathcal{S}b(.,f) \sim \mathcal{N}(0, \sigma_b^2). \tag{6.2}$$

6.2 Théorie pour les mélanges de gaussiennes

Nous allons étendre le filtrage de Wiener avec des mélanges de lois gaussiennes, comme loi *a priori* des sources. Dans un premier temps, nous développons les estimateurs sur des modèles de mélange de gaussiennes (MMG), puis nous verrons que l'extension à des modèles HMM est directe.

6.2.1 Estimateurs pour les mélanges de gaussiennes

Dans la suite de cette section, les modèles *a priori* des sources sont des mélanges de gaussiennes centrées. Bijaoui utilise déjà ces modèles pour opérer un filtrage de Wiener dans le cadre du débruitage [Bij02].

Nous considérons que les trames de chaque source $\mathcal{S}s_i(t,f)$, avec $i=1,2$, sont générées par un modèle de mélange de gaussiennes centrées, dont les paramètres sont les matrices de covariance diagonales $\sigma^2_{k,i}(f)$ où $k=1,\ldots,K_i$ et les poids des gaussiennes $\omega_{k,i}$. K_i est le nombre de composantes gaussiennes dans le modèle de la sources s_i.
On a donc les modèles *a priori* suivants :

$$p(\{\mathcal{S}s_1(.,f)\}_f) \approx \sum_{k=1}^{k_1} \omega_{k,1} \frac{1}{(2\pi)^{N/2}\prod_f \sigma_{k,1}(f)} \exp\left[-\frac{1}{2}\sum_f \frac{|\mathcal{S}s_1(.,f)|^2}{\sigma^2_{k,1}(f)}\right] \tag{6.3}$$

$$p(\{\mathcal{S}s_2(.,f)\}_f) \approx \sum_{k=1}^{k_2} \omega_{k,2} \frac{1}{(2\pi)^{N/2}\prod_f \sigma_{k,2}(f)} \exp\left[-\frac{1}{2}\sum_f \frac{|\mathcal{S}s_2(.,f)|^2}{\sigma^2_{k,2}(f)}\right], \tag{6.4}$$

où $\sum_{k=1}^{k_1}\omega_{k,1}=1$ et $\sum_{k=1}^{k_2}\omega_{k,2}=1$ (les poids sont normalisés). Notez que l'utilisation de la TFCT à fenêtre induit une approximation dans cette densité de probabilité *a priori*, du fait du fenêtrage.
On peut ré-écrire ces densités dans un formalisme de données incomplètes où les données non observées sont l'indexe de la composante gausssienne active k_i ($i=1,2$) dans chaque mélange.

$$p(\mathcal{S}s_1(t,f)|q_1(t)=k_1) \approx \frac{1}{(2\pi)^{N/2}\prod_f \sigma_{k_1,1}(f)} \exp\left[-\frac{1}{2}\sum_f \frac{|\mathcal{S}s_1(t,f)|^2}{\sigma^2_{k_1,1}(f)}\right] \tag{6.5}$$

$$p(q_1(t)=k_1) = \omega_{k_1,1}\,. \tag{6.6}$$

De même pour la source s_2 :

$$p(\mathcal{S}s_2(t,f)|q_2(t)=k_2) \approx \frac{1}{(2\pi)^{N/2}\prod_f \sigma_{k_2,2}(f)} \exp\left[-\frac{1}{2}\sum_f \frac{|\mathcal{S}s_2(t,f)|^2}{\sigma^2_{k_2,2}(f)}\right] \tag{6.7}$$

$$p(q_2(t)=k_2) = \omega_{k_2,2}\,. \tag{6.8}$$

Notez que dans ce modèle statistique, les réalisations des sources $\mathcal{S}s_i(t,f)$, pour une trame t et différentes fréquences f, ne sont pas indépendantes, contrairement au cas gaussien. Par contre, elles sont indépendantes conditionnellement à l'état $q_i(t)$.

Pour déduire les estimations des sources avec l'estimateur de l'espérance conditionnelle, nous allons sommer les estimations conditionnellement aux couples de composantes actives $(q_1(t), q_2(t))$ au temps t.

Notons $\gamma_{k_1,k_2}(t)$, la probabilité d'avoir le couple (k_1, k_2) comme composantes gaussiennes actives au temps t, conditionnellement aux observations, c'est-à-dire :

$$\gamma_{k_1,k_2}(t) = p\left(q_1(t) = k_1, q_2(t) = k_2 | \{\mathcal{S}x(t,f)\}_{f=1,\ldots,N}\right) \tag{6.9}$$

On a alors :

$$E(\mathcal{S}s_1(t,f)|x) = \sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2} E(\mathcal{S}s_1(t,f)|x,k_1,k_2)\gamma_{k_1,k_2}(t) \tag{6.10}$$

$$E(\mathcal{S}s_2(t,f)|x) = \sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2} E(\mathcal{S}s_2(t,f)|x,k_1,k_2)\gamma_{k_1,k_2}(t)\,. \tag{6.11}$$

En effet, on a par marginalisation :

$$p(\mathcal{S}s_i(t,f)|x) = \sum_{k_1}\sum_{k_2} p(\mathcal{S}s_i(t,f), q_1(t) = k_1, q_2(t) = k_2|x), \tag{6.12}$$

et, par la loi de Bayes :

$$p(\mathcal{S}s_i(t,f), q_1(t) = k_1, q_2(t) = k_2|x) = p(\mathcal{S}s_i(t,f)|x,k_1,k_2)\cdot p(k_1,k_2|\{\mathcal{S}x(t,f)\}_f), \tag{6.13}$$

d'où :

$$p(\mathcal{S}s_i(t,f)|x) = \sum_{k_1}\sum_{k_2} p(\mathcal{S}s_i(t,f)|x,k_1,k_2)\gamma_{k_1,k_2}(t), \tag{6.14}$$

car $\gamma_{k_1,k_2}(t) = p(k_1,k_2|\{\mathcal{S}x(t,f)\}_f)$. On en déduit les formules (6.10) et (6.11) en écrivant l'expression de l'espérance conditionnelle.

Il suffit donc pour estimer s_1 et s_2 avec l'espérance conditionnelle d'estimer ces sources conditionnellement au couple de composantes gaussiennes actives et de sommer en fonction des probabilités de ces couples conditionnellement aux observations.

Or, conditionnellement au couple de composantes gaussiennes actives (k_1, k_2), les densités *a priori* sont gaussiennes. On retombe donc sur un filtre de Wiener.

$$E(\mathcal{S}s_1(t,f)|x,k_1,k_2) = \frac{\sigma^2_{k_1,1}(f)}{\sigma^2_{k_1,1}(f)+\sigma^2_{k_2,2}(f)+\sigma^2_b}\mathcal{S}x(t,f) \tag{6.15}$$

$$E(\mathcal{S}s_2(t,f)|x,k_1,k_2) = \frac{\sigma^2_{k_2,2}(f)}{\sigma^2_{k_1,1}(f)+\sigma^2_{k_2,2}(f)+\sigma^2_b}\mathcal{S}x(t,f)\,. \tag{6.16}$$

On en déduit le filtrage suivant :

$$E(\mathcal{S}s_1(t,f)|x) = \sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}\gamma_{k_1,k_2}(t)\frac{\sigma^2_{k_1,1}(f)}{\sigma^2_{k_1,1}(f)+\sigma^2_{k_2,2}(f)+\sigma^2_b}\mathcal{S}x(t,f) \quad (6.17)$$

$$E(\mathcal{S}s_2(t,f)|x) = \sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}\gamma_{k_1,k_2}(t)\frac{\sigma^2_{k_2,2}(f)}{\sigma^2_{k_1,1}(f)+\sigma^2_{k_2,2}(f)+\sigma^2_b}\mathcal{S}x(t,f)\,. \quad (6.18)$$

Il reste donc à estimer les poids des différents filtres de Wiener dans l'estimation, $\gamma_{k_1,k_2}(t)$.

Conditionnellement aux composantes actives (k_1,k_2), les densités *a priori* des sources sont des gaussiennes centrées de matrices de covariance respectives $\sigma^2_{k_1,1}(f)$ et $\sigma^2_{k_2,2}(f)$. Comme $\mathcal{S}x(t,f)=\mathcal{S}s_1(t,f)+\mathcal{S}s_2(t,f)+\mathcal{S}b(t,f)$, la densité de $\mathcal{S}x(t,f)$ conditionnellement au couple (k_1,k_2) est aussi une gausienne centrée de matrice de covariance $\sigma^2_{k_1,1}(f)+\sigma^2_{k_2,2}(f)+\sigma^2_b$ (comme produit de convolution des trois gaussiennes).
On en déduit la vraisemblance suivante :

$$p(\mathcal{S}x(t,f)|k_1,k_2) \approx \frac{1}{(2\pi)^{N/2}\prod_f\sqrt{\sigma^2_{k_1,1}(f)+\sigma^2_{k_2,2}(f)+\sigma^2_b}}\exp\left[-\frac{1}{2}\sum_f\frac{|\mathcal{S}x(t,f)|^2}{\sigma^2_{k_1,1}(f)+\sigma^2_{k_2,2}(f)+\sigma^2_b}\right] \quad (6.19)$$

D'où la probabilité *a posteriori* :

$$\gamma_{k_1,k_2}(t) \propto \frac{\omega_{k_1,1}\omega_{k_2,2}}{\prod_f\sqrt{\sigma^2_{k_1,1}(f)+\sigma^2_{k_2,2}(f)+\sigma^2_b}}\exp\left[-\frac{1}{2}\sum_f\frac{|\mathcal{S}x(t,f)|^2}{\sigma^2_{k_1,1}(f)+\sigma^2_{k_2,2}(f)+\sigma^2_b}\right], \quad (6.20)$$

avec $\sum_{k_1,k_2}\gamma_{k_1,k_2}(t)=1$.

D'où finalement l'algorithme suivant pour l'estimation de sources avec l'estimateur de l'espérance conditionnelle :

Algorithme 1

estimation des probabilités conditionnelles

$$\gamma_{k_1,k_2}(t) \propto \frac{\omega_{k_1,1}\omega_{k_2,2}}{\prod_f\sqrt{\sigma^2_{k_1,1}(f)+\sigma^2_{k_2,2}(f)+\sigma^2_b}}\exp\left[-\frac{1}{2}\sum_f\frac{|\mathcal{S}x(t,f)|^2}{\sigma^2_{k_1,1}(f)+\sigma^2_{k_2,2}(f)+\sigma^2_b}\right], \quad (6.21)$$

avec $\sum_{k_1,k_2}\gamma_{k_1,k_2}(t)=1$.

filtrage

$$E_t(\mathcal{S}s_1|x) = \sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}\gamma_{k_1,k_2}(t)\frac{\sigma^2_{k_1,1}(f)}{\sigma^2_{k_1,1}(f)+\sigma^2_{k_2,2}(f)+\sigma^2_b}\mathcal{S}x(t,f) \quad (6.22)$$

$$E_t(\mathcal{S}s_2|x) = \sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}\gamma_{k_1,k_2}(t)\frac{\sigma^2_{k_2,2}(f)}{\sigma^2_{k_1,1}(f)+\sigma^2_{k_2,2}(f)+\sigma^2_b}\mathcal{S}x(t,f)\,. \quad (6.23)$$

On voit donc que l'on filtre la TFCT du signal observé avec tous les filtres de Wiener associés aux couples de DSP et on pondère ces filtres par la probabilité du couple d'états associé $\gamma_{k_1,k_2}(t)$ qui dépend elle-même du signal observé. Il s'agit donc là d'un filtrage temps-fréquence adaptatif.

6.2.2 Extension aux Modèles de Markov Cachés

Dans le cas de Modèles de Markov cachés à densité conditionnelle gaussienne, les poids *a priori* des gaussiennes au temps t dépendent des états observés aux temps $t-1,\ldots,t-K$ où K est l'ordre du HMM.

De manière simple, on peut étendre le filtrage précédemment calculé aux HMM comme densités à priori des sources, car la fonction $\gamma_{k_1,k_2}(t)$ peut être estimée dans ce cas par l'algorithme forward-backward [EM02].

6.3 Estimation des paramètres des modèles de source sonore

Dans une phase d'apprentissage, les paramètres des modèles GMM des sources doivent être estimés sur des exemples typiques de ces sources. Contrairement au cas gaussien, cette estimation est plus délicate.

6.3.1 Méthodes utilisées pour l'estimation des paramètres

En effet, en ce qui concerne les covariances $\sigma_{k_i,i}(f)$ des composantes gaussiennes, il existe un algorithme permettant de les estimer au maximum de vraisemblance : l'algorithme EM. Néanmoins, cet algorithme permet seulement d'atteindre un maximum local, qui dépend fortement de l'initialisation. Cette étape d'initialisation est donc cruciale pour une bonne estimation.

Nous avons opté pour l'initialisation des covariances par un algorithme de type quantification vectoriel. Les données considérées sont le carré du module de la TFCT des signaux sources $|\mathcal{S}s(t,f)|^2$ (on peut aussi prendre le logarithme). On commence par initialiser un premier cluster en calculant la moyenne globale des données. A chaque étape, on sépare chaque cluster en deux perpendiculairement à la première direction principale de la matrice de covariance des données appartenant au cluster, c'est-à-dire la direction de plus grand étalement du cluster. Par un algorithme de plus proches voisins, on obtient alors deux nouveaux clusters, dont on calcule les centroïdes. Finalement, cet algorithme permet d'obtenir un nombre 2^p de régions dans l'espace du spectre de puissance. Le calcul des variances correspondantes ou

DSP locales est alors direct.

Cependant un autre problème difficile concernant l'estimation des modèles concerne le choix du nombre de composantes gaussiennes K_i. Remarquons qu'il existe des méthodes fondées sur un compromis entre la qualité du modèle, mesurée par la vraisemblance, et la complexité du modèle. Cependant ces méthodes nécessitent généralement le réglage d'un hyperparamètre concernant la pondération entre qualité du modèle et complexité. Pour notre part, nous sommes contenté de tester les modèles avec différents nombres de composantes gaussiennes et de les comparer.

6.3.2 Estimation au maximum de vraisemblance des covariances : le problème de la dégénérescence

L'estimation au maximum de vraisemblance des DSP (covariances diagonales) des sources, dans le modèle MMG, peut poser des problèmes, car la vraisemblance n'est pas bornée. Ceci est un problème bien connu et une manière de lever cette dégénérescence consiste à ajouter une loi *a priori* inverse gamma sur les variances en jeu. De cette manière, la loi *a posteriori* est bornée [SMD01] et on évite d'avoir des variances qui tendent vers zéro.

6.4 Conclusion provisoire

Nous avons établi des formules d'estimation des sources dans le cadre de l'estimateur de l'espérance conditionnelle dans le cadre bayésien de mélange de sources avec un seul capteur. Ces formules étendent celles de Wiener au cas de mélanges de lois gausssiennes.

Nous avons traité le problème de l'estimation des paramètres des modèles de sources sonores qui se fait dans une phase d'apprentissage préalable, sur des exemples séparés.

Une des limitations du modèle proposé ici est que chaque son est modélisé par une composante gaussienne donc par une DSP, y compris en ce qui concerne l'intensité. Il doit donc y avoir théoriquement autant de DSP que d'intensités possibles pour une note donnée, ce qui en pratique n'est pas acceptable. Nous allons voir dans la troisième partie un autre modèle dans lequel le facteur d'amplitude (intensité) est un paramètre séparé, à part entière, de façon à modéliser réellement un son par DSP indépendamment de l'intensité. Chaque DSP modélisera alors une forme spectrale ("spectral shape"), en particulier sans faire intervenir son amplitude.

Chapitre 7

Utilisation du logarithme du spectre de puissance

Nous allons voir dans ce chapitre comment transposer le modèle de séparation de sources avec des MMG lorsqu'on utilise le logarithme du spectre de puissance. L'utilisation du logarithme du module de la TFCT est classique en traitement de la parole et se justifie d'un point de vue psychophysique.

7.1 Logarithme du module d'une variable aléatoire gaussienne

Nous étudions ici une variable aléatoire gaussienne centrée monodimensionelle de variance σ^2.

On note donc $y \sim \mathcal{N}(0, \sigma)$ et on étudie $m(\sigma) = E(\log|y|)$, l'espérance du logarithme de cette v.a. en fonction de la variance.

Nous avons :

$$\begin{aligned} m(\sigma) &= E(\log|y|) && (7.1)\\ &= \int_y \log|y| \frac{\exp\left[-\frac{|y|^2}{2\sigma^2}\right]}{\sqrt{2\pi}\sigma} dy && (7.2)\\ &= \log\sigma + \int_y \log\left|\frac{y}{\sigma}\right| \frac{\exp\left[-\frac{|y|^2}{2\sigma^2}\right]}{\sqrt{2\pi}\sigma} dy && (7.3)\\ &= \log\sigma + \int_x \log|x| \frac{\exp\left[-\frac{|x|^2}{2}\right]}{\sqrt{2\pi}} dx && (7.4)\\ &= \log\sigma + E(\log|x|), && (7.5) \end{aligned}$$

où $x \sim \mathcal{N}(0, 1)$.

Par simulation, nous avons évalué :

$$m(\sigma) \approx \log \sigma - 0.635. \tag{7.6}$$

De même, pour déduire la variance du logarithme d'une variable gaussienne monodimensionelle, on étudie $\beta^2(\sigma) = E[(\log|y| - m(\sigma))^2]$.
Les calculs sont les suivants :

$$\begin{aligned}
\beta^2(\sigma) &= E[(\log|y| - m(\sigma))^2], &(7.7)\\
&\approx E[(\log|y| - \log\sigma + 0.635)^2], &(7.8)\\
&\approx E[(\log\left|\frac{y}{\sigma}\right| + 0.635)^2], &(7.9)\\
&\approx E[(\log|x| + 0.635)^2], &(7.10)
\end{aligned}$$

où $x \sim \mathcal{N}(0,1)$.

Par simulation, nous avons trouvé que $\beta^2(\sigma) \approx 1.23$ et la variance du logarithme est indépendante de la variance du processus gaussien. Cette propriété de constance est partagée par tous les moments centrés d'ordre supérieur ou égal à deux.

7.2 Modèle multi-gaussien pour le logarithme du spectre de puissance

Dans la phase de séparation avec l'extension du filtre de Wiener aux modèles MMG, les probabilités $\gamma_{k_1,k_2}(t)$ d'activation du couple de composantes k_1, k_2 au temps t peuvent être estimées sur le logarithme du spectre de la TFCT du mélange $\log|\mathcal{S}x(t,f)|$. Nous cherchons donc la loi qui lie ce logarithme à la fonction $\gamma_{k_1,k_2}(t)$, en considérant le modèle de base sur la TFCT.
Comme nous souhaitons rester dans le cadre d'une modélisation MMG, nous supposerons que cette loi conditionnelle est gaussienne, dont nous cherchons à évaluer la moyenne $m_{k_1,k_2}(f)$ et la variance $\beta^2_{k_1,k_2}(f)$.

$$\gamma_{k_1,k_2}(t) \propto \frac{\omega_{k_1,1}\omega_{k_2,2}}{\prod_f \beta_{k_1,k_2}(f)} \exp\left[-\sum_f \frac{\left|\log|\mathcal{S}x(t,f)| - m_{k_1,k_2}(f)\right|^2}{2\beta^2_{k_1,k_2}(f)}\right] \tag{7.11}$$

$$\text{avec } \sum_{k_1}\sum_{k_2}\gamma_{k_1,k_2}(t) = 1\,. \tag{7.12}$$

A partir de la relation (7.6), on obtient :

$$\begin{aligned}
\sigma^2_{k_1,k_2}(f) &= \sigma^2_{k_1}(f) + \sigma^2_{k_2}(f), &(7.13)\\
m_{k_1,k_2}(f) &= \frac{1}{2}\log[\sigma^2_{k_1,k_2}(f)] - 0.635, &(7.14)
\end{aligned}$$

on en déduit la formule suivante :

$$m_{k_1,k_2}(f) = \frac{1}{2}\log[\sigma^2_{k_1}(f) + \sigma^2_{k_2}(f)] - 0.635, \tag{7.15}$$

et de plus, comme :

$$2m_{k_i}(f) = \log[\sigma^2_{k_i}(f)] - 2 \times 0.635, \tag{7.16}$$

On déduit la relation suivante, entre moyennes sur les logarithmes de la TFCT :

$$m_{k_1,k_2}(f) = \frac{1}{2}\log[\exp(2m_{k_1}(f)) + \exp(2m_{k_2}(f))], \tag{7.17}$$

où $m_{k_i}(f)$ est la moyenne du logarithme du spectre de la source s_i, calculée pour la composante k_i du MMG dans la phase d'apprentissage.
Notons que cette formule est compatible avec celle donnée par Roweis [Row00] :

$$m_{k_1,k_2}(f) = \max\left[m_{k_1}(f), m_{k_2}(f)\right], \tag{7.18}$$

qui semble être une approximation de (7.17), car le fait de prendre le logarithme de la somme des exponentiels de deux quantités revient approximativement à prendre le maximum de ces deux quantités. La formule (7.17) correspond en fait à une version lissée et dérivable de la fonction maximum de deux quantités.

En fait, les deux formules correspondent de manière plus générale à :

$$m_{k_1,k_2}(f) = \frac{1}{\delta}\log[\exp(\delta \cdot m_{k_1}(f)) + \exp(\delta \cdot m_{k_2}(f))], \tag{7.19}$$

avec dans le cas du maximum, $\delta \to \infty$ et dans notre cas $\delta = 2$.
De même, on obtient $\beta^2_{k_1,k_2}(f) = 1.23$, quels que soient k_1 et k_2.

Rappelons cependant que cette propriété de constance des moments d'ordre supérieur ou égal à deux, est intrinsèque à la modélisation dans le domaine de la TFCT. En général, en traitement de la parole, on raisonne différemment, car on modélise directement le logarithme du spectre de puissance (ou plus souvent le cepstre) par un modèle MMG, sans autre hypothèse sur le signal lui-même. Dans notre raisonnement, au contraire, on part d'une modélisation sur le signal, ou plutôt sur sa transformée de Fourier à court terme, puis on en déduit un modèle pour le logarithme de la TFCT. Cela explique que les relations énoncées ici ne soient pas classiques en traitement de la parole.

7.3 Utilisation de densités asymétriques dans les MMG

Comme on peut le voir sur la figure 7.1, la modélisation du logarithme de la valeur absolue du spectre par un modèle gaussien n'est pas très exacte, sous l'hypothèse de gaussiannité du

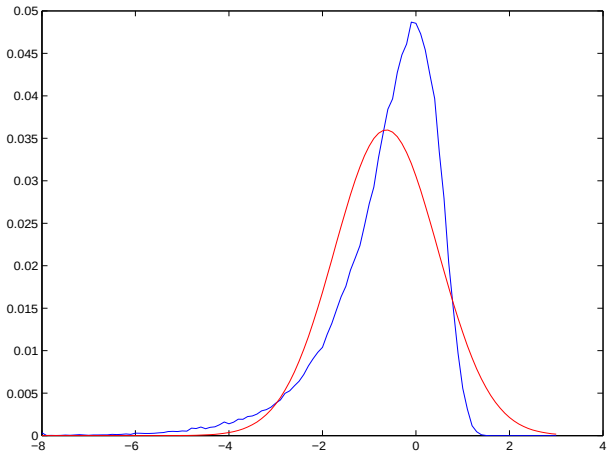

FIG. 7.1 – Distribution du logarithme du module d'une v.a. gaussienne centrée de variance 1 (en bleu) et modélisation gaussienne (en rouge)

processus sous-jacent. On peut donc essayer de modéliser chaque composante des modèles MMG, portant sur le logarithme du spectre, par une densité asymétrique décomposée comme un mélange de gaussiennes dont les moyennes sont liées et les variances sont fixées à l'avance.

La figure 7.2 indique en rouge la forme de la densité asymétrique choisie avec un mélange de quatre gaussiennes.

Dans l'étape de maximisation de l'algorithme EM, à l'étape k, le calcul de la nouvelle valeur du paramètre, $m_i^{(k+1)}(f)$ pour la composante i du mélange, en fonction des données $\log|\mathcal{S}x(t,f)|$ et des probabilités d'activation de chacune des gaussiennes $\gamma_{i,j}^k(t)$, est donc modifiée :

$$m_i^{(k+1)}(f) = \frac{\sum_t \sum_{j=1}^p \frac{\gamma_{i,j}^k(t)}{\sigma_j^2}(\log|\mathcal{S}x(t,f)| - m_j)}{\sum_t \sum_{j=1}^p \frac{\gamma_{i,j}^k(t)}{\sigma_j^2}}, \tag{7.20}$$

où i est l'indice de l'état du MMG et j est l'indice de la gaussienne dans le sous-MMG qui modélise la densité asymétrique. m_j et σ_j sont des paramètres fixés à l'avance.

Remarque sur la dégénérescence du critère du maximum de vraisemblance

Avec la modélisation sur le logarithme du module du spectrogramme des signaux, on a montré que les densités conditionnelles du MMG ont des variances fixes. Il n'y a donc plus de problème de dégénérescence du critère de maximum de vraisemblance. Cela s'explique en

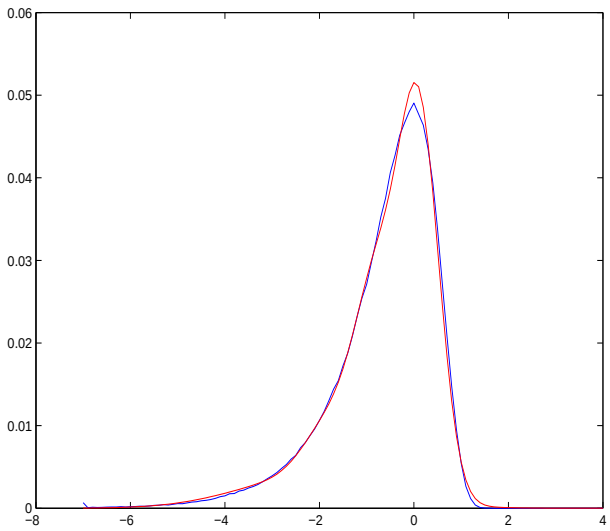

FIG. 7.2 – Distribution du logarithme du module d'une v.a. gaussienne centrée de variance 1 (en bleu) et modélisation asymétrique par un mélange de quatre gaussiennes (en rouge)

pratique par le fait que l'on prend comme variable $\log(|\mathcal{S}x(t,f)| + \epsilon)$, où ϵ est une petite constante.

7.4 Conclusion

Nous avons étudié l'utilisation du logarithme du module de la Transformée de Fourier à Court Terme des signaux sonores dans le cadre du modèle de séparation de sources avec des modèles *a priori* MMG, présentés au chapitre précédent.

Nous avons montré que les moments d'ordre supérieur ou égal à deux d'un modèle MMG portant sur le logarithme du spectre des signaux sont constants conditionnellement à la composante gaussienne. Ceci rend inutile l'estimation des variances dans ce modèle.

Troisième partie

Modèles à facteur d'amplitude

Cette partie est consacrée aux modèles à facteur d'amplitude dans le cadre de la séparation de sources. L'utilisation de facteurs d'amplitude permet de modéliser des phénomènes non-stationnaires à partir de statistiques d'ordre deux.

Dans le premier chapitre, nous exposons le modèle gaussien à facteur d'amplitude, qui est le modèle de base pour appréhender les phénomènes non-stationnaires.

Dans le chapitre suivant, nous décrivons deux modèles composites à facteurs d'amplitude et leur utilisation pour la séparation de sources : le modèle de mélange de gaussiennes à facteurs d'amplitude (MMGA) et le modèle à base de dictionnaire de DSP.

Le troisième chapitre est consacré à l'estimation ou à l'intégration des facteurs d'amplitude dans le cadre de la séparation de sources avec un seul capteur.

Un quatrième chapitre expose le problème de l'apprentissage de dictionnaires de DSP pour le deuxième type de modèle.

Chapitre 8

Modélisation de signaux localement stationnaires

8.1 Introduction

L'utilisation du filtre de Wiener et de ses dérivés pose problème dans le sens que l'on estime dans la phase d'apprentissage des DSP qui sont absolues et qui ne tiennent pas compte du fait que la même forme spectrale peut être présente avec des intensités différentes, du fait de la non-stationnarité.

De même, si la puissance globale des signaux d'apprentissage et des signaux présents dans le mélange (les sources à estimer) ne correspondent pas, la séparation ne sera pas possible. Autrement dit, il faut connaître *a priori* la puissance des sources dans le mélange, ou, dans le cas du débruitage, le rapport signal à bruit. En fait, on aimerait bien estimer l'énergie locale liée à un processus gaussien sous-jacent, donc à une DSP, pour chaque trame.

L'utilisation de modèles non-stationnaires dont la non-stationnarité est appréhendée par un facteur d'amplitude ou d'échelle a fait l'objet d'études "anciennes" [AM74] et a connu un regain d'intérêt récent, notamment en séparation de sources non-stationnaires [PSS00, PC01] et en débruitage [PSWS01].

8.2 Modèles à facteur d'amplitude

On utilisera ainsi dans la suite des modèles gaussiens à facteur d'amplitude (ou activation). Ce sont des modèles de la forme :

$$s(t) = \sqrt{a(t)} \times b(t) \tag{8.1}$$

où $b(t)$ est le processus gaussien sous-jacent et $a(t)$ est un facteur non négatif variant lentement par rapport à la taille de la trame d'analyse pour la TFCT (Stationnarité locale). La figure 8.1 illustre ce concept. Dans le domaine temps-fréquence, cela donne :

$$\mathcal{S}s(t,f) = \sqrt{a(t)} \times \mathcal{S}b(t,f) \tag{8.2}$$

Le modèle probabiliste correspondant est donné par la relation :

$$p(\mathcal{S}s(t,f)|a(t)) \approx \frac{1}{\sqrt{\prod_f 2\pi a(t)\sigma^2(f)}} \exp\left[-\frac{1}{2}\sum_f \frac{|\mathcal{S}x(t,f)|^2}{a(t)\sigma^2(f)}\right], \tag{8.3}$$

où $\sigma(f)$ est la DSP du processus gaussien sous-jacent. Notons que l'on peut utiliser une densité *a priori* pour le paramètre d'amplitude $a(t)$, comme par exemple une loi inverse gamma [LP00].

Il y a deux façons de combiner des modèles gaussiens à facteurs d'amplitude pour obtenir un modèle réaliste de sources :

- Soit on considère qu'un seul modèle gaussien est actif en même temps (combinaison par un "ou exclusif"), alors on tombe sur des modèles de type Modèles de Mélanges de Gaussiennes à facteur d'Amplitude (MMGA) que nous développerons au chapitre suivant.
- Soit on considère que tous ces modèles gaussiens à facteur d'amplitude s'additionnent pour former la source vraie (combinaison par un "et"), et on retombe sur un formalisme proche de l'analyse en composantes indépendantes. On a désigné ces modèles "Modèles à base de dictionnaires de DSP" du fait du caractère additif des "sous-sources" sous-jacentes. Nous développerons ce modèle par la suite.

Le chapitre suivant est consacré au modèle MMGA et au modèle à base de dictionnaire de DSP.

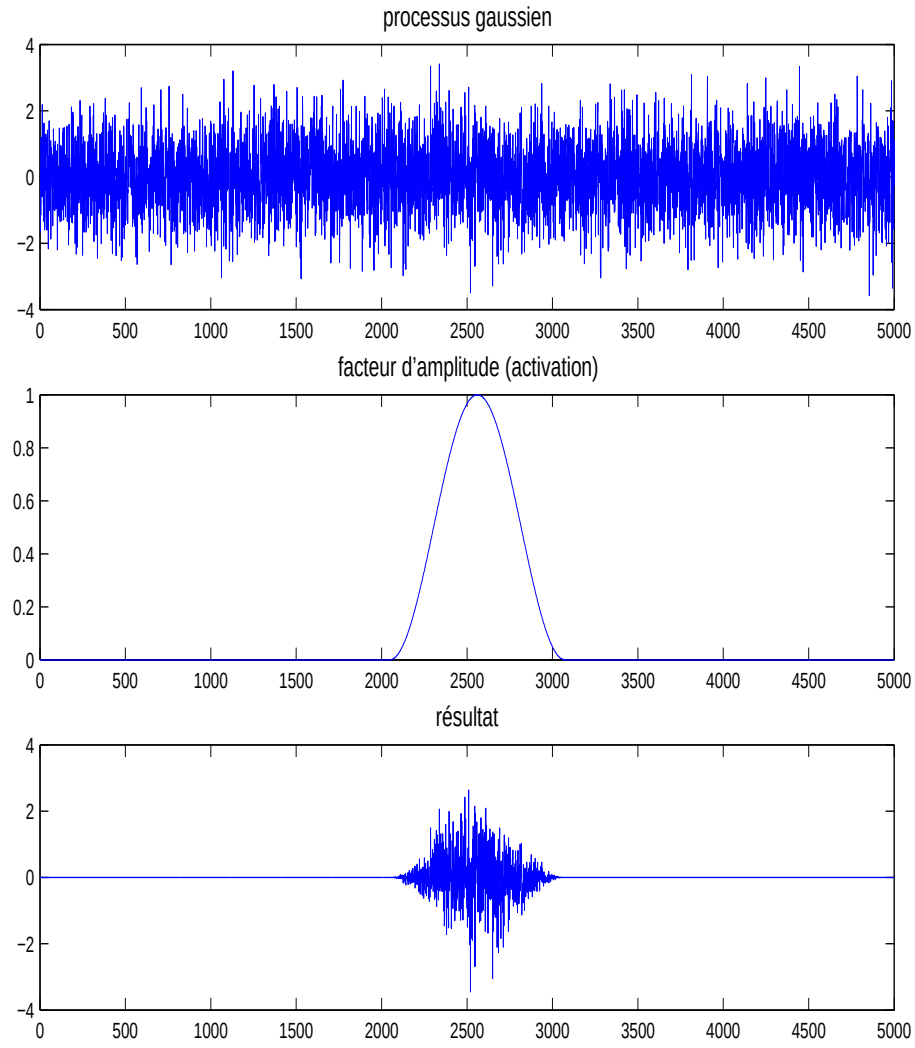

FIG. 8.1 – Processus gaussien modulé par un facteur d'amplitude

Chapitre 9

Applications pour la séparation de sources avec un capteur

Nous proposons ici deux modèles dans lequel les DSP sont définies à un facteur d'amplitude près. Cela induit de nouvelles variables dans le modèle dont nous aborderons l'estimation ou l'intégration dans le chapitre suivant. Ces modèles correspondent au modèle *gaussian scaled model* introduit dans [AM74] dans le cas d'une gaussienne unique. Dans le cas d'une combinaison des modèles gaussiens à facteur d'amplitude avec un "ou" exclusif, cela étend les modèles de mélange de gaussiennes, devenant ainsi des modèles de mélange de gaussiennes à facteur d'amplitude (MMGA). Nous abordons cette modélisation dans la première partie du chapitre et étudions ses propriétés pour l'inférence des sources dans le cadre de notre problème de séparation.

Dans le cas d'une combinaison des modèles gaussiens à facteur d'amplitude avec un "et", on obtient des modèles à base de dictionnaires de DSP. Nous établissons une formule de Wiener généralisée pour ces modèles dans la seconde partie de ce chapitre.

9.1 Le modèle MMGA

Nous supposons encore que l'équation de mélange est de la forme :

$$\mathcal{S}x(t,f) = \mathcal{S}s_1(t,f) + \mathcal{S}s_2(t,f) + \mathcal{S}b(t,f), \tag{9.1}$$

où b est un bruit blanc gaussien de variance σ_b^2.

Nous allons définir une loi *a priori* sur les sources à partir d'un modèle génératif qui sera une extension des modèles de mélange de gaussiennes avec des facteurs d'amplitude ou d'échelle.

Nous voulons ajouter à chaque DSP ou variance $\sigma^2_{k_i,i}(f)$ de la composante gaussienne k_i, pour la source $i = 1, 2$, un paramètre d'amplitude $a_{k_i,i}$, de façon à ce que la DSP observée soit $a_{k_i,i} \cdot \sigma^2_{k_i,i}(f)$ et que le paramètre $a_{k_i,i}$ soit évalué indépendamment de la DSP.

Pour cela, nous utilisons une gaussienne à facteur d'amplitude (gaussian scaled model, en Anglais). Nous définissons ainsi une variable aléatoire intermédiare $z_{k_i,i}(t,f) = \sqrt{a_{k_i,i}(t)} b_{k_i,i}(t,f)$ où $b_{k_i,i} \sim \mathcal{N}(0, \sigma^2_{k_i,i}(f))$ est une v.a. gaussienne et $a_{k_i,i}$ est une autre variable aléatoire non négative, ayant éventuellement une densité à priori $p_{k_i,i}(a_{k_i,i})$ [PSWS01].

La densité de la variable (conditionnelle) observée $z_{k_i,i}(t,f)$, à t fixé, est :

$$p\left(z_{k_i,i}(t,f)\right) = \int_{a_{k_i,i}} \frac{1}{(2\pi a_{k_i,i})^{N/2} \prod_f \sigma_{k_i,i}(f)} \exp\left[-\sum_f \frac{|z_{k_i,i}(t,f)|^2}{2 a_{k_i,i} \sigma^2_{k_i,i}(f)} \right] p_{k_i,i}(a_{k_i,i}) da_{k_i,i} \; . \quad (9.2)$$

La densité des observations complètes $z_{k_i,i}(t,f), a_{k_i,i}(t)$ est

$$p(z_{k_i,i}(t,f), a_{k_i,i}(t) | \sigma_{k_i,i}) = \frac{1}{(2\pi a_{k_i,i}(t))^{N/2} \prod_f \sigma_{k_i,i}(f)} \exp\left[-\sum_f \frac{|zk_i,i(t,f)|^2}{2 a_{k_i,i}(t) \sigma^2_{k_i,i}(f)} \right] p_{k_i,i}(a_{k_i,i}(t)) \quad (9.3)$$

A partir de ce modèle de gaussienne à facteur d'amplitude, on construit un modèle de mélange de gaussiennes à facteur d'amplitude (MMGA)

$$\begin{aligned} p(\mathcal{S}s_i(t,f) | q_i(t) = k_i, a_{k_i,i}) &= \frac{1}{(2\pi a_{k_i,i})^{N/2} \prod_f \sigma_{k_i,i}(f)} \exp\left[-\sum_f \frac{|\mathcal{S}s_i(t,f)|^2}{2 a_{k_i,i} \sigma^2_{k_i,i}(f)} \right] & (9.4) \\ p(a_{k_i,i} | q_i(t) = k_i) &= p_{k_i,i}(a_{k_i,i}) & (9.5) \\ p(q_i(t) = k_i) &= \omega_{k_i,i} \; . & (9.6) \end{aligned}$$

9.1.1 Théorie pour les estimateurs des sources

Conditionnellement au couple d'états (k_1, k_2) et au couple d'amplitudes correspondant $(a_{k_1,1}, a_{k_2,2})$, l'estimation suivant l'espérance conditionnelle est un filtre de Wiener de la forme :

$$\begin{aligned} E(\mathcal{S}s_1(t,f) | k_1, k_2, a_{k_1,1}, a_{k_2,2}) &= \frac{a_{k_1,1} \sigma^2_{k_1,1}(f)}{a_{k_1,1} \sigma^2_{k_1,1}(f) + a_{k_2,2} \sigma^2_{k_2,2}(f) + \sigma^2_b} \mathcal{S}x(t,f), & (9.7) \\ E(\mathcal{S}s_2(t,f) | k_1, k_2, a_{k_1,1}, a_{k_2,2}) &= \frac{a_{k_2,2} \sigma^2_{k_2,2}(f)}{a_{k_1,1} \sigma^2_{k_1,1}(f) + a_{k_2,2} \sigma^2_{k_2,2}(f) + \sigma^2_b} \mathcal{S}x(t,f), & (9.8) \end{aligned}$$

puisque les sources sont alors conditionnellement gaussiennes.

De même, conditionnellement aux amplitudes $a_{k_1,1}$ et $a_{k_2,2}$ associées aux composantes gaussiennes k_1 et k_2 de chaque modèle MMGA, on a la probabilité *a posteriori* $\gamma_{k_1,k_2,a_{k_1,1},a_{k_2,2}}(t)$ des composantes actives k_1 et k_2, conditionnellement aux observations et au jeu d'amplitudes

$\{a_{l_1,1}, a_{l_2,2}\}$:

$$\gamma_{k_1,k_2,a_{k_1,1},a_{k_2,2}}(t) \propto \omega_{k_1,1}\omega_{k_2,2}\frac{\exp\left[-\frac{1}{2}\sum_f \frac{|\mathcal{S}x(t,f)|^2}{a_{k_1,1}\sigma^2_{k_1,1}(f)+a_{k_2,2}\sigma^2_{k_2,2}(f)+\sigma_b^2}\right]}{\prod_f \sqrt{\left|a_{k_1,1}\sigma^2_{k_1,1}(f) + a_{k_2,2}\sigma^2_{k_2,2}(f) + \sigma_b^2\right|}}. \tag{9.9}$$

On a fait le même raisonnement que dans le cas des modèles MMG, c'est-à-dire que les sources sont conditionnellement gaussiennes (lorsque le couple d'états est fixé et les facteurs d'amplitude aussi. On en déduit que l'observation est également gaussienne comme somme de deux variables aléatoires gaussiennes.
Conditionnellement aux facteurs d'amplitude, on a donc les estimées suivantes :

$$E(\mathcal{S}s_1(t,f)|\{a_{k_1,1}, a_{k_2,2}\}) = \sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2} \gamma_{k_1,k_2,a_{k_1,1},a_{k_2,2}}(t)\frac{a_{k_1,1}\sigma^2_{k_1,1}(f)}{a_{k_1,1}\sigma^2_{k_1,1}(f) + a_{k_2,2}\sigma^2_{k_2,2}(f) + \sigma_b^2}\mathcal{S}x(t,f) \tag{9.10}$$

$$E(\mathcal{S}s_2(t,f)|\{a_{k_1,1}, a_{k_2,2}\}) = \sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2} \gamma_{k_1,k_2,a_{k_1,1},a_{k_2,2}}(t)\frac{a_{k_2,2}\sigma^2_{k_2,2}(f)}{a_{k_1,1}\sigma^2_{k_1,1}(f) + a_{k_2,2}\sigma^2_{k_2,2}(f) + \sigma_b^2}\mathcal{S}x(t,f) \tag{9.11}$$

9.1.2 Interprétation du modèle à facteur d'amplitude

Une variable aléatoire gaussienne à facteur d'amplitude correspond, de façon imagée, à un instrument qui jouerait toujours la même note de façon continue et dont l'intensité serait modifiée 'lentement' (amplitude constante à l'échelle d'une trame) par un opérateur. Le modèle de mélange de gaussienne à facteur d'amplitude implique par rapport à notre exemple qu'un seul instrument joue à la fois. Nous verrons par la suite qu'un autre modèle connexe est possible, dans lequel tous les instruments sont présents à la fois, modulo les facteurs d'amplitude.

9.2 Modèles à base de dictionnaires de DSP

9.2.1 Théorie

Nous supposons que nous disposons de deux sous-dictionnaires $\{\sigma_k^2(f)\}_{k\in Q_1}$ et $\{\sigma_k^2(f)\}_{k\in Q_2}$ caractéristiques de deux sources s_1 et s_2. Q_1 et Q_2 sont les ensembles d'indices des deux sous-dictionnaires, avec $Q_1 \cap Q_2 = \oslash$. L'idée est de réunir les deux sous-dictionnaires en un dictionnaire $\{\sigma_k^2(f)\}_{k\in Q_1\cup Q_2}$ sur lequel on décompose le spectre du signal composite $x = s_1 + s_2 + b$. On a donc :

$$|\mathcal{S}x(t,f)|^2 \approx \sum_{k\in Q_1\cup Q_2} a_k(t)\sigma_k^2(f) + \sigma_b^2, \tag{9.12}$$

où les coefficients $a_k(t)$ sont à calculer pour chaque indice de temps t. On va montrer que l'on a alors la formule d'estimation des sources suivantes (filtre de Wiener généralisé) :

$$\widehat{\mathcal{S}s_1}(t,f) = \frac{\sum_{k\in Q_1} a_k(t)\sigma_k^2(f)}{\sum_{k\in Q_1\cup Q_2} a_k(t)\sigma_k^2(f)+\sigma_b^2}\mathcal{S}x(t,f), \tag{9.13}$$

$$\widehat{\mathcal{S}s_2}(t,f) = \frac{\sum_{k\in Q_2} a_k(t)\sigma_k^2(f)}{\sum_{k\in Q_1\cup Q_2} a_k(t)\sigma_k^2(f)+\sigma_b^2}\mathcal{S}x(t,f)\,. \tag{9.14}$$

9.2.2 Expression du filtrage bayésien

Afin d'établir les formules (9.13) et (9.14), on peut considérer que l'on a affaire à $|Q_1|+|Q_2|$ sources indépendantes additives dont on possède une Densité Spectrale de Puissance $\sigma_k^2(f)$ et qui sont regroupées en fonction de leur appartenance à l'un ou l'autre des sous-dictionnaires. Pour une source gaussienne $s_j(t)$ de DSP $\sigma_j^2(f)$ et de facteur d'amplitude $a_j(t)$, on a la formule de Wiener suivante :

$$\widehat{\mathcal{S}s_j}(t,f) = \frac{a_j(t)\sigma_j^2(f)}{\sum_{k\in Q_1\cup Q_2} a_k(t)\sigma_k^2(f)+\sigma_b^2}\mathcal{S}x(t,f)\,. \tag{9.15}$$

Le modèle sous-jacent est $\mathcal{S}x(t,f) = \sum_{k\in Q_1\cup Q_2}\mathcal{S}s_k(t,f)$ et le regroupement des sources suivant $\mathcal{S}s_1(t,f) = \sum_{k\in Q_1}\mathcal{S}s_k(t,f)$ et $\mathcal{S}s_2(t,f) = \sum_{k\in Q_2}\mathcal{S}s_k(t,f)$ permettent d'établir les formules (9.13) et (9.14).
Si l'on explicite le modèle de source utilisé, on a :

$$\mathcal{S}x(t,f) = \underbrace{\sum_{k\in Q_1}\sqrt{a_k(t)}\cdot\mathcal{S}b_k(t,f)}_{\mathcal{S}s_1(t,f)} + \underbrace{\sum_{k\in Q_2}\sqrt{a_k(t)}\cdot\mathcal{S}b_k(t,f)}_{\mathcal{S}s_2(t,f)} + \mathcal{S}b(t,f), \tag{9.16}$$

où $\mathcal{S}b_k(t,f)$ est un processus gaussien centré de matrice de covariance $\sigma_k^2(f)$.

9.3 Conclusion

Nous avons considéré deux modèles de sources différents à partir du modèle gaussien à facteur d'amplitude : le modèle MMGA et le modèle à base de dictionnaire de DSP.

En ce qui concerne la complexité des modèles pour la séparation de sources, le nombre de paramètres à estimer pour le modèle MMGA est en $O(Q^s)$, où Q est le nombre de DSP par source et s est le nombre de source. Pour le modèle à base de dictionnaire de DSP, la complexité est en $O(Q\cdot s)$, ce qui est très avantageux.

Les facteurs d'amplitude sont des hyperparamètres des modèles de sources. Nous allons traiter au chapitre suivant de la question de leur estimation ou de leur intégration.

Chapitre 10

Estimation ou intégration des facteurs d'amplitude

10.1 Introduction

Nous avons vu dans le chapitre précédent comment étendre le filtre de Wiener dans le cas de modèles à facteur d'amplitude. Il reste néanmoins, au préalable, à expliciter l'estimation de ces facteurs d'amplitude. Ceux-ci vont être estimés au maximum de vraisemblance.

Nous explicitons ici le cas des dictionnaires de DSP, mais le cas des mélanges de gaussiennes est tout à fait similaire.

10.2 Estimation des facteurs d'amplitude

On considère le modèle bruité $\mathcal{S}x(t,f) = \sum_{k\in Q_1\cup Q_2} \sqrt{a_k(t)}\mathcal{S}b_k(t,f) + \mathcal{S}b(t,f)$. Comme $\mathcal{S}b_k(t,f)$ sont des processus gaussiens indépendants, $\mathcal{S}x(t)$ est aussi un processus gaussien conditionnellement aux coefficients $a_k(t)$ et les variances s'ajoutent.
D'où la vraisemblance suivante :

$$p(\mathcal{S}x(t,f)|\ldots,a_k(t),\ldots) \approx \frac{1}{\prod_f \sqrt{2\pi \sum_{k\in Q_1\cup Q_2} a_k(t)\sigma_k^2(f)+\sigma_b^2}} \exp\left[-\sum_f \frac{|\mathcal{S}x(t,f)|^2}{2\sum_{k\in Q_1\cup Q_2} a_k(t)\sigma_k^2(f)+\sigma_b^2}\right] \quad (10.1$$

On en déduit l'estimation au maximum de vraisemblance des coefficients $a_k(t)$ sous contrainte de positivité en annulant la dérivée du logarithme de l'expression précédente :

$$\forall k \in Q_1\cup Q_2,\ \forall t,\ \sum_f \sigma_k^2(f) \frac{\sum_{k\in Q_1\cup Q_2} a_k(t)\sigma_k^2(f)+\sigma_b^2-|\mathcal{S}x(t,f)|^2}{\left(\sum_{k\in Q_1\cup Q_2} a_k(t)\sigma_k^2(f)+\sigma_b^2\right)^2} = 0, \qquad (10.2)$$

sous la contrainte $\forall k \in Q_1\cup Q_2,\ \forall t,\ a_k(t) \geq 0$.

10.2.1 Algorithme de décompositions non-négatives

L'équation précédente peut se transformer en un schéma itératif dans lequel le dénominateur est gardé constant à chaque étape [LS00].
A l'étape l, on a les estimées $a_k^{(l)}(t)$ des coefficients d'amplitude. On approxime le dénominateur de l'équation (10.2) par :

$$D^{(l)}(t,f) = \left(\sum_{k \in Q_1 \cup Q_2} a_k^{(l)}(t) \sigma_k^2(f) + \sigma_b^2 \right)^2 . \tag{10.3}$$

L'équation d'estimation des coefficients peut alors se ré-écrire pour cette étape :

$$a_k^{(l+1)}(t) = \arg\min_{a_k(t)} \sum_f \frac{\left| |\mathcal{S}x(t,f)|^2 - \sum_{k \in Q_1 \cup Q_2} a_k(t) \sigma_k^2(f) - \sigma_b^2 \right|^2}{D^{(l)}(t,f)}, \tag{10.4}$$

sous contrainte de positivité des $a_k(t)$.
On voit que cette équation s'apparente à une approximation aux moindres carrés pondérés du spectre du mélange par la DSP résultant des différentes composantes du dictionnaire.
Pour résoudre cette minimisation sous contrainte, on introduit le Lagrangien

$$L^{(l)}(t, a_k(t), \lambda_k(t)) = \sum_f \frac{\left| |\mathcal{S}x(t,f)|^2 - \sum_{k \in Q_1 \cup Q_2} a_k(t) \sigma_k^2(f) - \sigma_b^2 \right|^2}{D^{(l)}(t,f)} - \sum_{k \in Q_1 \cup Q_2} \lambda_k(t) \cdot a_k(t) \tag{10.5}$$

En dérivant le Lagrangien $L^{(l)}(t, a_k(t), \lambda_k(t))$ par rapport à $a_k(t)$, on obtient donc

$$\sum_f \sigma_k^2(f) \frac{|\mathcal{S}x(t,f)|^2 - \sum_{k \in Q_1 \cup Q_2} a_k(t) \sigma_k^2(f) - \sigma_b^2}{D^{(l)}(t,f)} = \lambda_k(t) . \tag{10.6}$$

La contrainte de positivité se traduit par l'équation [Ber99] :

$$\lambda_k(t) \cdot a_k(t) = 0 . \tag{10.7}$$

D'où :

$$\sum_f \sigma_k^2(f) \frac{|\mathcal{S}x(t,f)|^2 - \sum_{k \in Q_1 \cup Q_2} a_k(t) \sigma_k^2(f) - \sigma_b^2}{D^{(l)}(t,f)} \cdot a_k(t) = 0 . \tag{10.8}$$

On exploite là encore le schéma itératif :

$$a_k^{(l)}(t) \sum_f \sigma_k^2(f) \frac{|\mathcal{S}x(t,f)|^2}{D^{(l)}(t,f)} = a_k^{(l+1)}(t) \sum_f \sigma_k^2(f) \frac{\sum_{k \in Q_1 \cup Q_2} a_k^{(l)}(t) \sigma_k^2(f) + \sigma_b^2}{D^{(l)}(t,f)} . \tag{10.9}$$

Finalement, on obtient la formule d'estimation itérative suivante [LS00]

$$a_k^{(l+1)}(t) = a_k^{(l)}(t) \frac{\sum_f \sigma_k^2(f) \frac{|\mathcal{S}x(t,f)|^2}{D^{(l)}(t,f)}}{\sum_f \sigma_k^2(f) \frac{\sum_{k \in Q_1 \cup Q_2} a_k^{(l)}(t) \sigma_k^2(f) + \sigma_b^2}{D^{(l)}(t,f)}} . \tag{10.10}$$

D'où finalement l'algorithme 2 d'estimation au maximum de vraisemblance des coefficients $a_k(t)$

Algorithme 2

– Initialiser les coefficients $a_k^{(0)}(t)$

– A l'étape l

1 : Calculer le dénominateur

$$D^{(l)}(t,f) = \left(\sum_{k \in Q_1 \cup Q_2} a_k^{(l)}(t) \sigma_k^2(f) + \sigma_b^2 \right)^2 \tag{10.11}$$

2 : ré-estimer les coefficients

$$a_k^{(l+1)}(t) = a_k^{(l)}(t) \frac{\sum_f \sigma_k^2(f) \frac{|\mathcal{S}x(t,f)|^2}{D^{(l)}(t,f)}}{\sum_f \sigma_k^2(f) \frac{\sum_{k \in Q_1 \cup Q_2} a_k^{(l)}(t) \sigma_k^2(f) + \sigma_b^2}{D^{(l)}(t,f)}} \tag{10.12}$$

Décompositions parcimonieuses non-négatives

Afin d'améliorer les résultats de l'estimation des coefficients $a_k(t)$ et de remédier à la possibilité d'avoir d'éventuelles solutions multiples, puisque le problème d'estimation des coefficients d'amplitude au maximum de vraisemblance est à peu de chose près équivalent à une représentation non-négative du spectre du signal composite $|\mathcal{S}x(t,f)|^2$ sur un ensemble de DSP $\sigma_k^2(f)$. En effet, si le nombre de DSP est supérieur au nombre de bandes de fréquence, c'est à dire à la longueur de la fenêtre d'analyse de la TFCT, le système est sur-déterminé et il existe une infinité de solutions à ce problème. Au lieu d'estimer les facteurs d'amplitude au maximum de vraisemblance, nous proposons d'ajouter une loi *a priori* sur les facteurs d'amplitude $p(a_k)$, afin d'avoir une unique solution.

La loi *a posteriori* devient alors :

$$p(\ldots, a_k(t), \ldots | \mathcal{S}x(t,f)) \approx \frac{\exp\left[- \sum_f \frac{|\mathcal{S}x(t,f)|^2}{2 \sum_{k \in Q_1 \cup Q_2} a_k(t) \sigma_k^2(f) + \sigma_b^2} \right]}{\prod_f \sqrt{2\pi \sum_{k \in Q_1 \cup Q_2} a_k(t) \sigma_k^2(f) + \sigma_b^2}} \cdot \prod_{k \in Q_1 \cup Q_2} p(a_k(t)) \,. \tag{10.13}$$

Si on note $f(a_k) = -\log p(a_k)$, Lagrangien (10.5) devient :

$$L^{(l)}(t, a_k(t), \lambda_k(t)) = \sum_f \frac{\left| |\mathcal{S}x(t,f)|^2 - \sum_{k \in Q_1 \cup Q_2} a_k(t) \sigma_k^2(f) - \sigma_b^2 \right|^2}{D^{(l)}(t,f)} - \sum_{k \in Q_1 \cup Q_2} \lambda_k(t) \cdot a_k(t) + \sum_{k \in Q_1 \cup Q_2} f(a_k(t)) \tag{10.14}$$

On peut montrer que si le terme $\sum_k f(a_k)$ est Shur-concave, il en résulte une représentation parcimonieuse des $a_k(t)$ [KDRE99], c'est-à-dire que l'on va utiliser le moins de coefficients d'amplitude possible pour représenter le spectre, d'où il résultera une représentation efficace de ce spectre. En particulier, on s'intéressera aux fonctions de la forme $f(x) = \gamma x^\alpha$ (lois exponentielles), avec $\alpha < 1$ et où γ est un paramètre de parcimonie.

Du point de vue des calculs, peu de choses changent. On utilise le schéma suivant de mise à jour des coefficients (à la place de la formule (10.9)) :

$$a_k^{(l)}(t)\sum_f \sigma_k^2(f)\frac{|\mathcal{S}x(t,f)|^2}{D^{(l)}(t,f)} = a_k^{(l+1)}(t)\left(\sum_f \sigma_k^2(f)\frac{\sum_{k\in Q_1\cup Q_2} a_k^{(l)}(t)\sigma_k^2(f)+\sigma_b^2}{D^{(l)}(t,f)} + f'(a_k^{(l)}(t))\right) \quad (10.15)$$

Finalement, on obtient la formule d'estimation itérative suivante :

$$a_k^{(l+1)}(t) = a_k^{(l)}(t)\frac{\sum_f \sigma_k^2(f)\frac{|\mathcal{S}x(t,f)|^2}{D^{(l)}(t,f)}}{\sum_f \sigma_k^2(f)\frac{\sum_{k\in Q_1\cup Q_2} a_k^{(l)}(t)\sigma_k^2(f)+\sigma_b^2}{D^{(l)}(t,f)} + f'(a_k^{(l)}(t))}. \quad (10.16)$$

En particulier, dans le cas de la fonction $f(x) = \gamma x^\alpha$, on obtient :

$$a_k^{(l+1)}(t) = a_k^{(l)}(t)\frac{\sum_f a_k^{(l)}(t)\sigma_k^2(f)\frac{|\mathcal{S}x(t,f)|^2}{D^{(l)}(t,f)}}{\sum_f a_k^{(l)}(t)\sigma_k^2(f)\frac{\sum_{k\in Q_1\cup Q_2} a_k^{(l)}(t)\sigma_k^2(f)+\sigma_b^2}{D^{(l)}(t,f)} + \gamma'(a_k^{(l)}(t))^\alpha}. \quad (10.17)$$

Remarque : l'algorithme présenté au chapitre précédent avec les modèles de mélange de gaussiennes à facteur d'amplitude peut être vu comme un cas particulier de ce que nous présentons ici, avec en plus une contrainte sur le nombre de composantes non nulles, qui doit être égal à un dans chacun des sous-dictionnaires, c'est-à-dire à une seule DSP active pour chacune des sources.

10.2.2 Alternative pour l'estimation au maximum de vraisemblance : algorithme EM

Nous présentons ici une méthode alternative pour l'estimation des coefficients d'amplitude et des sources. Cette méthode est fondée sur l'algorithme EM de Bermond et Cardoso [BMC97, BC99a].

La méthode présentée ici est liée à la modélisation alternative :

$$\mathcal{S}x(t,f) = \sum_k b_k(t)\mathcal{S}s_k(t,f) + \mathcal{S}b(t,f), \quad (10.18)$$

où $s_k(t)$ sont des sources gaussiennes, centrées de matrice de covariance diagonale $\{\sigma_k^2(f)\}$ et $b(t)$ est un bruit blanc gaussien additif de variance σ_b^2.

Si on raisonne à t fixé (pour une trame donnée donc), on peut ré-écrire cette équation sous forme matricielle :

$$X = BS + Y_b, \quad (10.19)$$

où

- $X = [\mathcal{S}x(t,f_1),\ldots,\mathcal{S}x(t,f_n)]$,
- $B = [b_1(t),\ldots,b_Q(t)]$,
- $S = \begin{bmatrix} \mathcal{S}s_1(t,f_1) & \ldots & \mathcal{S}s_1(t,f_n) \\ \vdots & & \vdots \\ \mathcal{S}s_Q(t,f_1) & \ldots & \mathcal{S}s_Q(t,f_n) \end{bmatrix}$,
- $Y_b = [\mathcal{S}b(t,f_1),\ldots,\mathcal{S}b(t,f_n)]$.

On a posé $Q = |Q_1| + |Q_2|$.
On a donc à faire à un problème classique d'Analyse en Composantes Indépendantes ; à ceci près, que les sources et le mélange sont échantillonnés suivant l'axe fréquentiel, à t fixé, et non plus suivant l'axe temporel. De plus la matrice de mélange est "très" rectangulaire, puisque ses dimensions sont $1 \times K$. L'algorithme EM s'applique donc pour l'estimation conjointe des sources (espérance conditionnelle) et des paramètres d'amplitude (au maximum de vraisemblance).

D'après ce que l'on a vu dans la première partie du manuscrit (bibliographie sur l'Analyse en Composantes Indépendante dans le cas sous-déterminé), on aura pour l'estimation des coefficients :

$$B^{l+1} = R_{xs}^l {R_{ss}^l}^{-1} \tag{10.20}$$

$$\tag{10.21}$$

où l'on a dénoté les moments empiriques

$$R_{xs}^l = XE[S^T|X,B^l] \tag{10.22}$$

$$R_{ss}^l = E[SS^T|X,B^l], \tag{10.23}$$

Il reste à estimer ces deux espérances, ce qui ne pose pas de problèmes majeurs, car les lois *a priori* sur les sources sont gaussiennes.
En reprenant l'écriture scalaire, écrivons la loi des données complètes :

$$-\log p(\mathcal{S}x(t,f)|b_k,\mathcal{S}s_k(t,f)) \approx \sum_f \left[\frac{|\mathcal{S}x(t,f) - \sum_k b_k \mathcal{S}s_k(t,f)|^2}{2\sigma_b^2} + \sum_k \frac{|\mathcal{S}s_k(t,f)|^2}{2\sigma_k^2(f)} \right] \tag{10.24}$$

L'expression de gauche est quadratique en $\mathcal{S}s_k(t,f)$, on en déduit :

$$E(\mathcal{S}s_k(t,f)|\{b_j\},\mathcal{S}x(t,f)) = \frac{b_k\sigma_k^2(f)}{\sum_j b_j^2\sigma_j^2(f) + \sigma_b^2}\mathcal{S}x(t,f). \tag{10.25}$$

De même, on calcule la covariance de $\mathcal{S}s_k(t,f)$, en fonction de la fréquence f, pour une trame t fixée (Hessienne) :

$$H(f) = \begin{pmatrix} \frac{1}{\sigma_1^2(f)} & 0 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & \dots & 0 & \frac{1}{\sigma_Q^2(f)} \end{pmatrix} + \frac{1}{\sigma_b^2} \begin{pmatrix} b_1^2 & \dots & b_1 b_Q \\ \vdots & & \vdots \\ b_Q b_1 & \dots & b_Q b_Q \end{pmatrix}, \tag{10.26}$$

on obtient :

$$E(SS^T|X,B) = \sum_f H(f)^{-1} + E(S|X,B)E(S|X,B)^T \tag{10.27}$$

D'où finalement :

$$E(SS^T|X,B)_{i,j} = \sum_f \left[\delta_{i,j}\sigma_i^2(f) + \frac{|\mathcal{S}x(t,f)|^2 - (\sum_k b_k^2\sigma_k^2(f) + \sigma_b^2)}{(\sum_k b_k^2\sigma_k^2(f) + \sigma_b^2)^2} \cdot b_i\sigma_i^2(f) b_j \sigma_j^2(f) \right] \tag{10.28}$$

Il ne reste plus qu'à utiliser les formules (10.22) et (10.23) pour ré-estimer les amplitudes.

Notons que cette méthode fournit une estimée des sources gaussiennes s_k à chaque étape. Pour obtenir une estimation globale des sources s_1 et s_2, il suffit de sommer.
Par exemple :

$$\widehat{\mathcal{S}s_1}^l(t,f) = \sum_{k \in Q_1} b_k^l \cdot E(\mathcal{S}s_k(t,f)|\{b_j^l\}, \mathcal{S}x(t,f)) \tag{10.29}$$

10.2.3 Remarque sur le cas non bruité

Dans le cas non bruité ($\sigma_b \to 0$), la vraisemblance de coefficients d'amplitude :

$$p(\mathcal{S}x(t,f)|\dots, a_k(t), \dots) \approx \frac{\exp\left[-\sum_f \frac{|\mathcal{S}x(t,f)|^2}{2\sum_{k\in Q_1\cup Q_2} a_k(t)\sigma_k^2(f)} \right]}{\prod_f \sqrt{2\pi \sum_{k\in Q_1\cup Q_2} a_k(t)\sigma_k^2(f)}}, \tag{10.30}$$

peut ne pas être bornée. C'est ici encore un problème de dégénérescence (si $\mathcal{S}x(t,f) = 0$, pour une valeur de t et de f).

Une manière de résoudre ce problème consiste à ajouter une loi *a priori* sur la quantité $\sum_{k\in Q_1\cup Q_2} a_k(t)\sigma_k^2(f)$, sous forme de loi inverse gamma de paramètres $(\alpha, \frac{\beta}{2})$.

Rappelons que la densité d'une loi inverse gamma de paramètres $(\alpha, \frac{\beta}{2})$ est de la forme :

$$f(x) \propto \frac{1}{x^{\alpha+1}} \exp\left[-\frac{\beta}{2x} \right]. \tag{10.31}$$

La loi *a posteriori* devient :

$$p(\dots, a_k(t), \dots | \mathcal{S}x(t,f)) \propto \frac{\exp\left[-\frac{1}{2}\sum_f \frac{|\mathcal{S}x(t,f)|^2 + \beta}{\sum_{k\in Q_1\cup Q_2} a_k(t)\sigma_k^2(f)} \right]}{\prod_f \left(\sum_{k\in Q_1\cup Q_2} a_k(t)\sigma_k^2(f) \right)^{\alpha+1.5}}. \tag{10.32}$$

Cette loi *a posteriori* est alors bornée, pour toutes les valeurs de $a_k(t)$ et $\sigma_k(t)$.

10.3 Intégration des facteurs d'amplitude

Reprenons le modèle à facteur d'amplitude :

$$\begin{aligned} \mathcal{S}x(t,f) &= \sum_{k\in Q_1} \sqrt{a_k(t)}\cdot \mathcal{S}b_k(t,f) + \sum_{k\in Q_2} \sqrt{a_k(t)}\cdot \mathcal{S}b_k(t,f) + \mathcal{S}b(t,f), \quad (10.33)\\ &= \sum_{k\in Q_1\cup Q_2} \mathcal{S}s_k(t,f) + \mathcal{S}b(t,f), \quad (10.34)\end{aligned}$$

où l'on a posé $\mathcal{S}s_k(t,f) = \sqrt{a_k(t)}\cdot\mathcal{S}b_k(t,f)$, et $\mathcal{S}b(t,f)$ est un bruit blanc gaussien de variance σ^2.

La sous-source s_k est donc modélisée par un processus gaussien à facteur d'amplitude :

$$p(\mathcal{S}s_k(t,f)|a_k) \approx \frac{1}{\sqrt{\prod_f (2\pi a_k \sigma_k^2(f))}} \exp\left[-\frac{1}{2}\frac{|\mathcal{S}s_k(t,f)|^2}{a_k\sigma_k^2(f)}\right]. \quad (10.35)$$

On peut intégrer cette densité par rapport à a_k. Au préalable, on va donner une probabilité *a priori* pour a_k, afin d'éviter des problèmes numériques (si on intègre directement sans mettre une densité *a priori*, on obtient une densité sur s_k qui ne sera pas intégrable). Supposons que la densité *a priori* soit de la forme $a_k \sim \mathcal{IG}(\frac{1}{2}\nu, \frac{1}{2}\xi)$ (densité inverse gamma), où ν et ξ sont des hyperparamètres. Alors, on obtient [LP00] :

$$\begin{aligned} p(\mathcal{S}s_k(t,f)) &= \int_{a_k} p(\mathcal{S}s_k(t,f)|a_k)\cdot p(a_k)\cdot da_k \quad (10.36)\\ &\propto \frac{1}{\left(\xi + \left(\sum_f \frac{|\mathcal{S}s_k(t,f)|^2}{\sigma_k^2(f)}\right)\right)^{\frac{\nu+N}{2}}}, \quad (10.37)\end{aligned}$$

où N est la taille de la trame.

Cette nouvelle densité est appelé *t-distribution* de matrice de covariance $\{\sigma_k^2(f)\}$ avec ν degrés de liberté.

On peut alors obtenir une estimation des sous-sources $\mathcal{S}s_k(t,f)$ au maximum *a posteriori*, en fonction de l'observation $\mathcal{S}x(t,f)$, en utilisant les relations (10.34) et (10.37). L'expression de la densité *a posteriori* est indépendante des facteurs d'amplitude et est de la forme :

$$\begin{aligned} -\log p(s_1,\ldots,s_Q|x) &= \sum_f \left[\frac{|\mathcal{S}x(t,f) - \sum_{k\in Q_1\cup Q_2}\mathcal{S}s_k(t,f)|^2}{2\sigma_b^2}\right] \quad (10.38)\\ &\quad + \frac{\nu+N}{2}\sum_{k\in Q_1\cup Q_2}\log\left[\xi + \left(\sum_f \frac{|\mathcal{S}s_k(t,f)|^2}{\sigma_k^2(f)}\right)\right] + \text{cte.} \quad (10.39)\end{aligned}$$

Remarquons que le deuxième terme l'équation (10.38) n'est pas convexe. Il faut donc faire attention pour la minimisation de cette fonctionnelle, car il peut y avoir plusieurs minima

locaux. Pour obtenir l'estimée MAP, il faut donc déjà partir d'une "bonne" solution, puis utiliser l'équation d'optimalité :

$$\frac{\sum_{j\in Q_1\cup Q_2}\mathcal{S}s_j(t,f)-\mathcal{S}x(t,f)}{\sigma_b^2}+\frac{\mathcal{S}s_k(t,f)}{\sigma_k^2(f)}\cdot\frac{\nu+N}{\xi+\left(\sum_f\frac{|\mathcal{S}s_k(t,f)|^2}{\sigma_k^2(f)}\right)}=0. \tag{10.40}$$

On peut ré-écrire cette équation sous la forme suivante :

$$\forall k,\ \frac{\sum_{j\in Q_1\cup Q_2}\mathcal{S}s_j(t,f)-\mathcal{S}x(t,f)}{\sigma_b^2}+\frac{\mathcal{S}s_k(t,f)}{\tilde{a}_k(t)\sigma_k^2(f)}=0, \tag{10.41}$$

où :

$$\tilde{a}_k(t)=\frac{\xi+\left(\sum_f\frac{|\mathcal{S}s_k(t,f)|^2}{\sigma_k^2(f)}\right)}{\nu+N}. \tag{10.42}$$

L'équation (10.41) peut encore se transformer en :

$$\mathcal{S}s_k(t,f)=\frac{\tilde{a}_k(t)\sigma_k^2(f)}{\sum_j\tilde{a}_j(t)\sigma_j^2(f)+\sigma_b^2}\mathcal{S}x(t,f). \tag{10.43}$$

On en déduit un algorithme d'estimation des sous sources $\mathcal{S}s_k(t,f)$, en calculant itérativement les paramètres $\tilde{a}_k(t)$ suivant la relation (10.42) et les sous sources suivant la relation (10.43).

10.4 Conclusion

Nous avons traité dans ce chapitre du problème des facteurs d'amplitude comme paramètres supplémentaires du modèle. Nous avons exposé deux possibilités : soit l'estimation au maximum de vraisemblance de ces coefficients, soit leur intégration. Dans les deux cas, nous avons proposé un ou plusieurs algorithmes pour résoudre le problème.

Dans le prochain chapitre, nous allons traiter de l'optimisation de dictionnaire de DSP dans la phase d'apprentissage.

Chapitre 11

Apprentissage des dictionnaires de formes spectrales

Nous allons maintenant discuter de l'apprentissage des DSP dans le cadre de la méthode de séparation de sources avec des dictionnaires de DSP.

11.1 Le problème

En principe, l'apprentissage devrait être différent de celui relatif aux modèles MMG, car le critère à optimiser n'est pas le même. Dans le cas d'une source simple, c'est-à-dire ne contenant qu'un seul instrument, on peut continuer à estimer les DSP par un algorithme EM, dans le cadre de la modélisation avec des MMG. Cela dit, dans le cas de sources plus complexes (notamment lorsqu'elles contiennent des accords), il conviendrait d'optimiser le sous-dictionnaire en maximisant la vraisemblance :

$$p(\mathcal{S}s_i(t,f)|\ldots,a_k(t),\ldots) \approx \frac{1}{\prod_f \sqrt{2\pi \sum_{k\in Q_i} a_k(t)\sigma_k^2(f)+\sigma_b^2}} \exp\left[-\sum_f \frac{|\mathcal{S}s_i(t,f)|^2}{2\sum_{k\in Q_i} a_k(t)\sigma_k^2(f)+\sigma_b^2}\right] \quad (11.1)$$

en fonction des paramètres recherchés $\sigma_k^2(f)$ (les DSP) et des facteurs d'amplitude $a_k(t)$ (éventuellement itérativement).

La différence avec le chapitre précédent, c'est que l'optimisation ne porte plus seulement sur les coefficients $a_k(t)$, comme dans la phase de séparation, mais aussi sur les DSP $\sigma_k^2(f)$, que l'on souhaite estimer dans une phase d'apprentissage.

11.2 Algorithme

Nous allons maximiser la vraisemblance (11.1) par rapport aux facteurs d'amplitudes, grâce à l'algorithme développé dans le chapitre précédent et par rapport aux DSP $\sigma_k^2(f)$, itérativement (algorithme de relaxation des variables à maximiser). Remarquer que l'on pourrait utiliser le même genre d'algorithme pour les DSP que pour les coefficients d'amplitude. Cependant, nous souhaitons que les DSP soient normalisées, i.e. $\sum_f \sigma_k^2(f) = 1$. C'est pourquoi nous optons pour un algorithme de gradient projeté pour la maximisation suivant les DSP ([Hoy02]). On obtient l'algorithme suivant :

Algorithme 3

– Initialiser les DSP, par un algorithme de quantification vectorielle par exemple.

– A l'étape l, on a estimé les DSP $\sigma_k^l(f)$ et les coefficients $a_k^l(t)$.

1. Calculer les coefficients :

$$D^l(t,f) = \sum_{k\in Q_i} a_k^l(t)(\sigma_k^l(f))^2 + \sigma_b^2, \tag{11.2}$$

$$a_k^{l+1}(t) = a_k^l(t) \cdot \frac{\sum_f (\sigma_k^l(f))^2 \frac{|\mathcal{S}s_i(t,f)|^2}{(D^l(t,f))^2}}{\sum_f (\sigma_k^l(f))^2 \frac{1}{D^l(t,f)}}. \tag{11.3}$$

2. Mise à jour des DSP :

$$D_2^l(t,f) = \sum_{k\in Q_i} a_k^{l+1}(t)(\sigma_k^l(f))^2 + \sigma_b^2, \tag{11.4}$$

$$(\beta_k^l(f))^2 = \max\left[(\sigma_k^l(f))^2 - \mu_l \sum_t a_k^l(t) \cdot \frac{D_2^l(t,f) - |\mathcal{S}s_i(t,f)|^2}{(D_2^l(t,f))^2}, 0\right], \tag{11.5}$$

$$(\sigma_k^{l+1}(f))^2 = \frac{(\beta_k^l(f))^2}{\sum_f (\beta_k^l(f))^2} \tag{11.6}$$

11.3 Discussion

La principale difficulté dans ce problème d'optimisation concerne les problèmes d'initialisation dus à la multi-modalité du critère. Ceci rend difficile l'obtention des solutions (c'est-à-dire des DSP) qui présentent une réelle amélioration par rapport à un calcul des DSP par segmentation (notamment à travers des modèles de mélange de lois). En fait, dans le cadre de la modélisation spectrale en DSP caractéristiques, cela demeure un problème ouvert.

Pour comprendre l'intérêt du problème posé en terme de modèles, on peut dire que les

modèles de mélanges de lois de type MMG induisent une segmentation du spectrogramme de la source d'apprentissage pour former les DSP caractéristiques. Dans le cadre d'une source complexe ou composite, c'est-à-dire formée d'une superposition de notes provenant soit du même instrument (accords) ou d'instruments différents, cette approche par segmentation peut paraître sous-optimale en terme d'économie de la représentation. En effet, il vaut mieux avoir une DSP par note possible, qu'une DSP par accord possible.

L'approche proposée dans ce chapitre permet théoriquement d'obtenir une représentation de type superposition de notes. Cependant l'existence de nombreux minima locaux fait que cette résolution en pratique est difficile.

Quatrième partie

Expérimentation et évaluation

Cette partie concerne les aspects expérimentaux pour la séparation de sources. Il s'agit d'évaluer les algorithmes proposés dans les parties précédentes.

Le premier chapitre concerne la définition des critères d'évaluation. Les deux critères proposés sont le rapport signal à interférence (SIR) et le rapport signal à artefact (SAR).

Le second chapitre explique quelques aspects pratiques d'implémentation, notamment en ce qui concerne la modélisation MMG et l'apprentissage par méthode EM.

Le troisième chapitre concerne l'évaluation elle-même, sur un extrait de Jazz.

Chapitre 12

Critères d'évaluation

12.1 Protocole expérimental

L'essentiel de l'étude expérimentale se fera sur un morceau de jazz dont nous possédons des enregistrements séparés de la batterie d'une part et du piano avec la contrebasse d'autre part. L'intérêt de ce morceau est de nous permettre de faire un apprentissage sur les sources séparées tout en travaillant sur un vrai mixage. Nous utiliserons, sauf précisions contraire, 45 secondes d'enregistrements pour la phase d'apprentissage et 15 secondes pour les tests. Notons que les parties d'apprentissage et de test sont choisies pour avoir le même contenu perceptuel, afin que l'apprentissage soit possible sur une taille réduite d'enregistrement ; les contenus des deux parties ne sont cependant pas strictement identiques, afin de simuler des conditions réelles de fonctionnement.

12.2 Critères d'évaluation

Afin de pouvoir comparer les performances des différents systèmes de séparation de sources qui seront présentés par la suite, nous nous proposons de définir deux critères d'évaluation de la séparation.

Nous supposons que les deux sources originales s_1 et s_2 sont connues et qu'elles sont décorrélées. On note $\hat{s}_1$ et $\hat{s}_2$ leur estimée fournie par l'algorithme.

Considérons la projection orthogonale de chacune des sources estimées sur l'espace vectoriel engendré par les sources vraies. On peut donc écrire :

$$\hat{s}_1 = \alpha_1 s_1 + \alpha_2 s_2 + n_1, \tag{12.1}$$

$$\hat{s}_2 = \beta_1 s_1 + \beta_2 s_2 + n_2, \tag{12.2}$$

où n_1 et n_2 sont les résidus de la décomposition sur les deux sources. Ces résidus sont orthogonaux aux sources. Ceci correspond à une décomposition des sources estimées comme combinaison linéaire des sources vraies. On pourrait affiner le modèle en considérant que les sources estimées résultent d'une combinaison non-linéaire des sources vraies, de la forme :

$$\hat{s}_1 = \alpha_1 f_1(s_1) + \alpha_2 f_2(s_2) + n_1, \tag{12.3}$$

$$\hat{s}_2 = \beta_1 f_1(s_1) + \beta_2 f_2(s_2) + n_2, \tag{12.4}$$

où f_1 et f_2 sont des fonctions à estimer.

Dans le cadre d'une décomposition linéaire des sources estimées, nous définissons alors le rapport source à interférence (en Anglais : Source to Interference Ratio ou SIR) comme le rapport en dB entre la composante de source réel dans l'estimation $\alpha_1 s_1$ (dans le cas de la première source $\hat{s}_1$) et de la composante d'interférence $\alpha_2 s_2$.

Nous définissons aussi le rapport source à artefact (en Anglais : Source to Artefact Ratio ou SAR) comme le rapport en dB entre la somme des composantes des sources $\alpha_1 s_1 + \alpha_2 s_2$ et la composante de bruit n_1, dans le cas de la première source estimée.

Finalement, on a :

$$\mathrm{SIR}_1 = 20\log_{10}\left|\frac{\alpha_1}{\alpha_2}\right|\frac{\|s_1\|}{\|s_2\|} \quad \mathrm{SAR}_1 = 20\log_{10}\frac{\|\alpha_1 s_1+\alpha_2 s_2\|}{\|n_1\|} \tag{12.5}$$

$$\mathrm{SIR}_2 = 20\log_{10}\left|\frac{\beta_2}{\beta_1}\right|\frac{\|s_2\|}{\|s_1\|} \quad \mathrm{SAR}_2 = 20\log_{10}\frac{\|\beta_1 s_1+\beta_2 s_2\|}{\|n_2\|} \tag{12.6}$$

Notons que le SIR est une mesure du résiduel de la source non souhaitée dans l'estimation de l'autre source. Le SAR correspond à la quantité de bruit qui a été généré lors de la séparation des sources suivant l'algorithme testé. Pour en savoir plus sur ces critères d'évaluation, voir l'article [GBVF03].

Ajoutons cependant que si les critères proposés ici semblent raisonnables, tout en apportant plus d'information qu'un simple SNR (Signal to Noise Ratio, en Français : Rapport signal à bruit), il existe des raisons soit de modifier ces critères, soit en tous cas d'en cerner les limites.

Ces limites sont principalement dues à l'optimisation d'une norme ℓ_2 sur les signaux, ce qui signifie que la correspondance entre les sources estimées et les sources réelles est globale, l'enveloppe des signaux jouant un rôle prépondérant. Une manière de remédier à cet effet est soit de moyenner les critères sur des portions des signaux (découper le signal en trame et moyenner les différents SIR et SAR sur les différentes trames obtenues), soit d'utiliser une autre norme que la norme ℓ_2, telle que par exemple la norme ℓ_1.

Malgré ces limitations, nous utiliserons les SIR et SAR comme critères d'évaluations dans la suite de cet exposé.

Chapitre 13

Evaluation sur un morceau de Jazz

Dans ce chapitre, nous rapportons les résultats de tests des algorithmes que nous avons proposés dans les chapitres précédents et les comparons au filtre de Wiener standard. Nous avons utilisé un exemple de test et d'apprentissage issu d'un standard de Jazz. Il serait nécessaire de tester les algorithmes de manière beaucoup plus large sur un ensemble suffisamment grand de signaux, ce que nous n'avons pas pu faire dans le cadre de cette thèse, faute de temps et de matériel sonore disponible. Cependant, l'exemple traité ici permet de se faire une idée des possibilités ouvertes par la théorie que nous avons développée et donne un premier aperçu expérimental.

13.1 Conditions expérimentales

13.1.1 Apprentissage

Nous disposons tout d'abord d'exemples d'apprentissage, d'une durée de 45 secondes.

Pour l'apprentissage des paramètres des modèles MMG et MMGA, nous avons utilisé l'algorithme EM avec des distributions asymétriques, sur le logarithme de la valeur absolue de la TFCT des signaux.

L'apprentissage par l'algorithme EM peut engendrer des difficultés, car il faut éviter la convergence vers des solutions aberrantes, à savoir de variance nulle, ces solutions correspondant à une vraisemblance infinie. Il est donc nécessaire de mettre en oeuvre une approche permettant de contrôler ces variances lors de l'apprentissage. Il existe plusieurs alternatives pour ce problème, par exemple, dans le cas de matrices de covariance diagonales, imposer une valeur minimale aux différentes variances en jeu.

De manière plus élégante, H. Snoussi [SMD01] a proposé d'ajouter une loi *a priori* sur les

covariances de gaussiennes, sous forme de loi de Wishart inverse. La loi *a posteriori* résultante est alors bornée, ce qui assure de ne pas converger vers une solution dégénérée.

Dans le cadre de nos expérimentations, nous nous sommes contentés d'ajouter un bruit blanc gaussien de faible amplitude (avec un rapport de -80dB par rapport au signal sonore). Dans le cadre de nos modèles de mélange de gaussiennes centrées, cela permet de s'assurer de ne pas avoir une solution dégénérée.

Notons que, *a posteriori*, l'influence de l'addition du bruit sur le résultat de la séparation est très faible. En effet, les résultats de séparation obtenus (de l'ordre de 10 dB de séparation) sont d'un tout ordre de grandeur que le bruit ajouté préalablement (-80 dB). Celui-ci peut donc être considéré comme négligeable, dans le cadre de nos expérimentations.

Dans le cas du modèle MMGA, nous estimons les variances (DSP) à partir des signaux normalisés en énergie, trame à trame, de façon à ne pas avoir à estimer de facteur d'amplitude lors de l'apprentissage.

Pour initialiser les modèles, avant les itérations de l'EM, on a utilisé un algorithme de quantification vectorielle.

13.1.2 Test

Pour former le signal de test (mélange), on utilise 15 secondes de chaque source, sur des extraits distincts de l'ensemble d'apprentissage. Nous ajoutons ces deux extraits des sources, qui sont synchrones, donc réalistes du point de vue du mixage, pour former le mélange.

13.2 Modèles de Mélange de Gaussiennes à facteurs d'amplitude

Dans cette section, nous comparons le modèle MMG au modèle MMGA, dans lequel les énergies locales sont estimées. Nous avons aussi fait figurer dans les tableaux de résultats 13.1 (SIR) et 13.2 (SAR), le cas du filtre de Wiener standard, qui peut être considéré comme un cas particulier du modèle MMG avec une gaussienne unique, par source.

La première constatation est qu'il y a une amélioration par rapport au filtre de Wiener standard, d'abord lorsqu'on rajoute des gaussiennes dans les modèles, et surtout lorsqu'on intègre un modèle non stationnaire avec les facteurs d'amplitude. Ceci est d'autant plus frappant sur la batterie qui est fortement non-stationnaire. D'autre part, les mauvais résultats en ce qui concerne le SAR (distorsion) sur la batterie peuvent s'expliquer par le fait que dans le mélange, la batterie est dans une proportion moindre que la source piano/contrebasse (afin

états	source	MMG	MMGA
Wiener	piano	14.1	13.8
Wiener	drums	2.8	13.0
4	piano	17.4	14.6
4	drums	5.7	12.4
8	piano	17.4	14.9
8	drums	8.3	13.7

TAB. 13.1 – SIR pour chaque source en fonction du nombre de composantes dans le modèle

états	source	MMG	MMGA
Wiener	piano	20.1	23.2
Wiener	drums	-1.8	-3.2
4	piano	16.0	21.2
4	drums	-0.5	-2.2
8	piano	17.0	20.5
8	drums	-0.2	-2.2

TAB. 13.2 – SAR pour chaque source en fonction du nombre de composantes dans le modèle

d'avoir un mélange réaliste perceptuellement). Néanmoins, ce défaut est peu audible et c'est surtout la qualité de l'atténuation sur la batterie qui est remarquable à l'écoute.

13.3 Dictionnaires de DSP

13.3.1 Résultats

On présente dans les tableaux 13.3 et 13.4, les SIR et SAR pour le modèle MMGA (pour la comparaison), pour le modèle par dictionnaire de DSP avec apprentissage classique (EM/GMM) et avec apprentissage par optimisation du dictionnaire (en utilisant la norme ℓ_1 pour la régularisation).

13.3.2 Commentaires

La première constatation est que, sur ces exemples, les résultats par les deux types de méthodes (MMG ou dictionnaires) sont proches et il est difficile de dire quelle méthode est plus valable. Cependant il est important de noter que la méthode présentée dans cette section (par dictionnaire) est de complexité algorithmique bien moindre que la méthode MMGA. En

états	source	MMGA	dico	dico opt.
4	piano	14.6	13.9	17.5
4	drums	12.4	15.3	13.6
8	piano	14.9	15.9	16.9
8	drums	13.7	19.6	18.3

TAB. 13.3 – SIR pour chaque source en fonction du nombre de composantes dans le modèle

états	source	MMGA	dico	dico opt.
4	piano	21.2	23.0	17.6
4	drums	-2.2	-2.9	-0.5
8	piano	20.5	19.7	18.9
8	drums	-2.2	-1.1	-0.3

TAB. 13.4 – SAR pour chaque source en fonction du nombre de composantes dans le modèle

effet, pour p sources et Q DSP par source, la complexité de la méthode à base de modèle MMGA est en $O(Q^p)$, alors qu'avec la méthode présentée ici on est en $O(Q \cdot p)$.

Par ailleurs, il semble que l'optimisation du dictionnaire donne des résultats légèrement supérieurs à un apprentissage des DSP par segmentation (GMM).

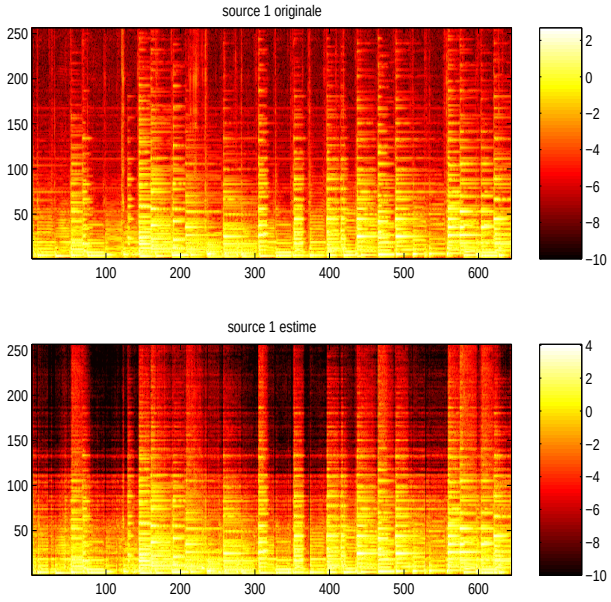

FIG. 13.1 – Spectrogrammes de la source piano/contrebasse originale et estimée

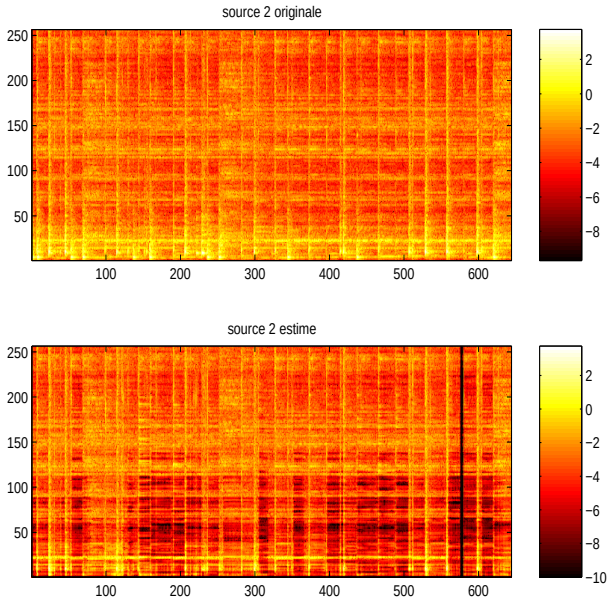

FIG. 13.2 – Spectrogrammes de la source de batterie originale et estimée

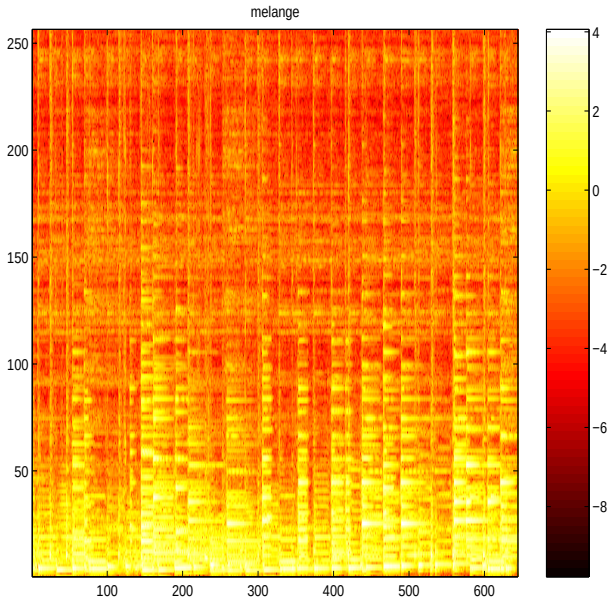

FIG. 13.3 – Spectrogramme du mélange

13.4 Conclusion

Dans ce chapitre, nous avons évalué les méthodes de séparation de sources avec un seul capteur qui ont été présentées dans les parties précédentes.

Sur cet exemple particulier, on observe une amélioration sensible par rapport à un filtrage de Wiener standard, du fait de la modélisation multi-gaussienne et aussi de l'intégration des facteurs d'amplitude.

D'autre part, la méthode à base de dictionnaires de DSP semble donner des résultats similaires à la méthode fondée sur le modèle MMGA. Ceci est encourageant car la méthode à base de dictionnaires est beaucoup moins coûteuse algorithmiquement, notamment dans le cas d'un nombre important de sources.

Il est naturellement essentiel de consolider ces premiers résultats expérimentaux par une batterie de tests exhaustifs qui pourront être conduits par exemple sur la base des données en cours de développement par l'Action Jeune Chercheur "séparation de sources" du GDR ISIS.

Cinquième partie

Conclusion et perspectives

Quelques problèmes restant à résoudre

Nous avons développé au cours de ce manuscrit des modèles de sources sonores pour la séparation avec un mélange unique et un modèle additif de mélange des sources.

Les modèles de source sont probabilistes et sont invariants au changement de phase. Ceci est un point faible par rapport à la modélisation de signaux audio, car il existe des relations plus ou moins déterministes entre les phases des différentes composantes fréquentielles de certains extraits audio, notamment dans le cas de signaux harmoniques. De même, la modélisation du mélange comme une simple addition des sources est trop simpliste par rapport à une prise de son ou un mixage réels. Nous développons ici quelques pistes à explorer pour pallier ces limitations dans les algorithmes développés lors de nos travaux.

Estimation de la phase

Nous ne développerons pas dans ce chapitre de modèle précis de phase pour les signaux audio, mais donnerons quelques idées de base qui pourrait servir de point de départ pour obtenir de meilleurs modèles des signaux sonores.

Les modèles développés dans cette thèse, à base de distributions gaussiennes centrées, sont insensibles à la phase. Or, les relations de phase à l'intérieur d'une trame et la continuité de la phase d'une trame à l'autre (de la TFCT du signal) sont perceptivement importantes dans les signaux sonores. En effet, si l'on reconstruit un signal analysé de trames pour lesquelles la cohérence des phases n'est pas assurée (au sein des trames et entre trames), le signal reconstruit présentera des défauts tout à fait perceptibles en particulier sur les parties harmoniques (ou voisées dans le cas de la parole).

Une piste serait de tenir compte de cette régularité de phase en introduisant des relations *a priori* sur les phases, dans le modèle de source sonore. Par exemple, les phases recalées de

trames adjacentes pour une même composante fréquentielle devraient être *a priori* proches. Ceci pourrait s'introduire sous forme de connaissances *a priori* dans l'approche bayésienne.

Raffinement des modèles de source sonore

Lors d'une prise de son réel ou d'un mixage, le signal sonore est modifié, par exemple à cause de la réverbération de la salle dans le cas d'une prise de son ou de l'utilisation d'effets sonores variés dans le cas d'un mixage. Une manière classique de tenir compte de ces effets sur la source sonore est d'introduire un filtre convolutif dans la modélisation. Ceci se traduit dans le domaine de Fourier par la multiplication de la TFCT de la source par la Transformée de Fourier du filtre. Le modèle peut alors s'écrire, dans le cadre du modèle par dictionnaire de DSP :

$$\mathcal{S}s_i(t,f) = H_i(f) \cdot \left[\sum_{k \in K_i} a_k(t) \mathcal{S}b_k(t,f) \right], \tag{13.1}$$

où $H_i(f)$ est la transformée de Fourier du filtre de convolution propre à la source $s_i(t)$, et $b_k(t)$ est un processus gaussien centré de variance $\sigma_k^2(f)$.

Notons que $H_i(f)$ ne dépend que de la fréquence et est constant d'une trame à l'autre. Dans le cadre d'une modélisation plus fine, on pourrait considérer que ce filtre n'est constant qu'à l'échelle de quelques trames. Cela permettrait de tenir compte par exemple d'un mouvement de la source sonore dans l'espace dans le cas d'une prise de son.

Pour ce qui est de l'estimation de ces filtres (un par source), ils pourraient être estimés au maximum de vraisemblance, conjointement aux facteurs d'amplitude, selon une généralisation de l'approche proposée dans cette thèse.

Conclusion

Le problème de la séparation de deux sources sonores avec un seul microphone a été abordé dans ce manuscrit. Deux méthodes d'estimation des sources ont été développées théoriquement et étudiées sur un cas réel. Ces méthodes sont des extensions du filtrage fréquentiel de Wiener à des sources stationnaires à court terme. La première méthode est fondée sur une modélisation *a priori* des sources par modèle de mélange de lois (Mélange de Gaussiennes). La seconde utilise des dictionnaires de formes spectrales, caractéristiques des sources. Ces deux méthodes sont connexes et même si la seconde méthode paraît algorithmiquement plus séduisante, il nous a paru intéressant d'exposer ces deux approches.

Si diverses méthodes pour le problème du capteur unique émergent, sans que le lien entre ces différentes approches soit toujours apparent, il nous semble que beaucoup de ces tentatives sont en réalité des extensions du filtrage de Wiener classique (ou de Kalman). C'est pourquoi nous avons voulu développer un cadre théorique (bayésien) autour des extensions possibles de ces méthodes de filtrage. Il faut noter que l'utilisation systématique de facteurs d'amplitude en plus des formes spectrales est, nous semble-t-il, une contribution importante de notre travail.

Peut-on pour autant dire que le problème est résolu ? Bien entendu, la réponse est non. Un des points faible, mais semble-t-il difficilement contournable, des méthodes de séparation de sources avec un seul microphone est la phase d'apprentissage. Dans un contexte réel, on peut difficilement espérer avoir des exemples de chaque instrument ou voie, pour l'apprentissage. D'autre part, nous avons étudié le problème dans le cas où les sources s'additionnent, c'est-à-dire d'un mixage basique. Or, dans le cadre d'utilisation d'effets plus ou moins complexes pour le mixage des sources ou de prise de son réelle, le problème de la séparation avec un seul capteur est un problème largement plus complexe.

Enfin, on peut encore souligner qu'une des limitations des modèles utilisés dans cette thèse est l'invariance à la phase. Intégrer un modèle de phase est à notre avis une perspective d'amélioration prioritaire.

Par ailleurs, les perspectives possibles de ces travaux sont nombreuses. Notamment l'intégration

de modèle perceptif (psychoacoustique) dans le filtrage ainsi que dans les critères d'évaluation est souhaitable. On peut aussi noter que le critère d'estimation des sources développé dans cette thèse est celui des moindres carrés (espérance conditionnelle), alors que les critères mesurés (SIR et SAR) sont différents. Il serait donc possible d'essayer d'optimiser directement les critères de séparation utilisés, avec un compromis nécessaire entre la distortion et les interférences.

D'autre part, dans le cadre des modèles à états, il serait possible de développer des modèles de motifs musicaux, c'est-à-dire étudier les séquences d'états à plus long terme. Ceci permettrait de mettre en oeuvre des *a priori* correspondant à des notions musicales telles que la tonalité ou le phrasé.

Pour finir, une autre perspective de travail est l'adaptation des modèles itérativement avec le processus de séparation, c'est-à-dire utiliser les sources estimées pour raffiner les modèles.

Annexe A

Calcul sur les facteurs d'amplitude (EM)

Dans le cadre de l'estimation de facteurs d'amplitude par méthode EM, on détaille ici le calcul de la covariance de S, pour une trame t fixée (partie III, chapitre 3) :

$$E(SS^T|X,B) = \sum_f H(f)^{-1} + E(S|X,B)E(S|X,B)^T, \tag{A.1}$$

où

$$H(f) = \begin{pmatrix} \frac{1}{\sigma_1^2(f)} & 0 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & \dots & 0 & \frac{1}{\sigma_Q^2(f)} \end{pmatrix} + \frac{1}{\sigma_b^2} \begin{pmatrix} b_1^2 & \dots & b_1 b_Q \\ \vdots & & \vdots \\ b_Q b_1 & \dots & b_Q b_Q \end{pmatrix}. \tag{A.2}$$

Pour inverser $H(f)$, on remarque que :

$$H^{-1}(f) = \left[I + \frac{1}{\sigma_b^2} \begin{pmatrix} \sigma_1^2(f) b_1^2 & \dots & \sigma_1^2(f) b_1 b_Q \\ \vdots & & \vdots \\ \sigma_Q^2(f) b_Q b_1 & \dots & \sigma_Q^2(f) b_Q b_Q \end{pmatrix} \right]^{-1} \cdot \begin{pmatrix} \sigma_1^2(f) & 0 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & \dots & 0 & \sigma_Q^2(f) \end{pmatrix} \tag{A.3}$$

La premier facteur de cette équation est de la forme $\left[I + xy^T\right]^{-1}$, où x et y sont des vecteurs. On utilise alors l'identité suivante :

$$\left[I + xy^T\right]^{-1} = I - \frac{1}{1 + \langle x, y \rangle} xy^T. \tag{A.4}$$

D'où, on obtient :

$$H^{-1}(f)_{i,j} = \delta_{i,j}\sigma_i^2(f) - \frac{b_i \sigma_i^2(f) b_j \sigma_j^2(f)}{\sum_k b_k^2 \sigma_k^2(f) + \sigma_b^2}. \tag{A.5}$$

D'où finalement :

$$E(SS^T|X,B)_{i,j} = \sum_f \left[\delta_{i,j}\sigma_i^2(f) + \frac{|\mathcal{S}x(t,f)|^2 - (\sum_k b_k^2\sigma_k^2(f) + \sigma_b^2)}{(\sum_k b_k^2\sigma_k^2(f) + \sigma_b^2)^2} \cdot b_i\sigma_i^2(f) b_j\sigma_j^2(f) \right], \quad \text{(A.6)}$$

car :

$$E(\mathcal{S}s_k(t,f)|\{b_j\}, \mathcal{S}x(t,f)) = \frac{b_k\sigma_k^2(f)}{\sum_j b_j^2\sigma_j^2(f) + \sigma_b^2}\mathcal{S}x(t,f). \quad \text{(A.7)}$$

Bibliographie

[AC97] S. Amari and J. Cardoso. Blind source separation — semiparametric statistical approach. *IEEE Transaction on Signal Processing*, 45(11) :2692–2700, December 1997.

[AM74] D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B*, 36 :99–102, 1974.

[Ama98] S. I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2) :251–276, 1998.

[BACEM97] Adel Belouchrani, Karim Abed Meraim, Jean-François Cardoso, and Éric Moulines. A blind source separation technique based on second order statistics. *IEEE Transactions on Signal Processing*, 45(2) :434–44, February 1997.

[BC95] Adel Belouchrani and Jean-François Cardoso. Maximum likelihood source separation by the expectation-maximization technique : deterministic and stochastic implementation. In *Proceedings of NOLTA*, pages 49–53, 1995.

[BC99a] Olivier Bermond and Jean-François Cardoso. Approximate likelihood for noisy mixtures. In *Proceedings of ICA '99, Aussois, France*, pages 325–330, 1999.

[BC99b] Olivier Bermond and Jean-François Cardoso. Méthodes de séparation de sources dans le cas sous-déterminé. In *Proceedings of GRETSI, Vannes, France*, pages 749–752, 1999.

[Ber99] D.P. Bertsekas. *Nonlinear Programming, second edition.* MIT, 1999.

[BGB01] L. Benaroya, R. Gribonval, and F. Bimbot. Représentations parcimonieuses pour la séparation de sources avec un seul capteur. In *Proceedings of the 18th Symposium GRETSI'01 on Signal and Image Processing, Toulouse*, 2001.

[BGB03] L. Benaroya, R. Gribonval, and F. Bimbot. Non negative sparse representation for wiener based source separation with a single sensor. In *Proceedings of ICASSP*, pages 613–616, Hong Kong, 2003.

[Bij02] A. Bijaoui. Wavelets, gaussian mixtures and wiener filtering. *Signal Processing*, 82 :709–712, 2002.

[BMC97] Olivier Bermond, Éric Moulines, and Jean-François Cardoso. Séparation et déconvolution aveugle de signaux bruités : modélisation par mélange de gaussiennes. In *Proceedings of GRETSI, Grenoble, France*, 1997.

[BS95] A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7 :1129–1159, 1995.

[Car98] J.F. Cardoso. Blind signal separation : statistical principles. In *Proceedings of IEEE*, volume 86, pages 2009–2025, 1998.

[Car01] Jean-François Cardoso. The three easy routes to independent component analysis ; contrasts and geometry. In *Proceedings of ICA 2001 workshop, San Diego*, 2001.

[Cas03] F. Castanié. *Analyse spectrale*. Hermès science, 2003.

[CDS98] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 1998.

[Coa98] Mark Coates. *Time-frequency Modelling*. PhD thesis, University of Cambridge, 1998.

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[CW00] M.A Casey and W. Westner. Separation of mixed audio sources by independent subspace analysis. In *Proceedings of the International Computer Music Conference, Berlin*, 2000.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 1977.

[Don95] D. L. Donoho. Denoising by soft-thresholding. *IEEE Transactions on Information Theory*, 41 :613–627, May 1995.

[Ell96] D. Ellis. *Prediction-driven computational auditory scene analysis*. PhD thesis, MIT Department of Electrical Engineering and Computer Science, 1996.

[EM02] Y. Ephraim and N. Merhav. Hidden markov processes. *IEEE Transactions on Information Theory*, 48(6) :1518–1569, June 2002.

[Eph92] Y. Ephraim. A bayesian estimation approach for speech enhancement using hidden markov models. *IEEE Transactions Signal Processing*, SP-40 :725–735, April 1992.

[GBVF03] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte. Proposals for performance measurement in source separation. In *Proceedings of ICA '03, Nara, Japan*, 2003.

[GJ95] Zoubin Ghahramani and Michael I. Jordan. Factorial hidden markov models. In *Advances in Neural Information Processing System*, volume 8, pages 472–478, 1995.

[HO00] A. Hyvärinen and E. Oja. Independent component analysis : Algorithms and applications. *Neural Networks*, 13 :411–430, 2000.

[Hoy02] P. O. Hoyer. Non-negative sparse coding. In *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing, Martigny, Switzerland*, pages 557–565, 2002.

[JLO03] Gil-Jin Jang, Te-Won Lee, and Yung-Hwan Oh. A subspace approach to single channel signal separation using maximum likelihood weighting filters. In *Proceedings of ICASSP*, 2003.

[KDRE99] K. Kreutz-Delgado, B.D. Rao, and K. Engan. Convex/schur-convex (csc) log-priors and sparse coding. In *Proceedings of the 6th Joint Symposium on Neural Computation*, 1999.

[LGBS98] Te-Won Lee, Mark Girolami, Anthony J. Bell, and Terrence J. Sejnowski. A unifying information-theoretic framework for independent component analysis, 1998.

[LP00] G. Mc Lachlan and D. Peel. *Finite mixture models*, chapter 7, pages 221–237. Wiley series in probability and statistics, 2000.

[LS00] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Proceedings of NIPS*, pages 556–562, 2000.

[MPZ98] S. Mallat, G. Papicolaou, and Z. Zhang. Adaptive covariance estimation of locally stationary processes. *Annals of Statistics*, 26(1) :1–47, 1998.

[PC01] D. T. Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of non stationary sources. *IEEE Transactions on Signal Processing*, 49(9) :1837–1848, 2001.

[PSS00] L. Parra, C. Spence, and P. Sajda. Higher-order statistical properties arising from the non-stationarity of natural signals. In *Advances in Neural Information Processing Systems*, volume 13, Denver, December 2000.

[PSWS01] J. Portilla, V. Strela, M.J. WainWright, and E. Simoncelli. Adaptive wiener denoising using a gaussian scale of mixture model in the wavelet domain. In *Proceedings of the 8th international onference on Image Processing*, Thessaloniki,Greece, October 2001.

[Rab89] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–285, 1989.

[Rob01] C.P. Robert. *The bayesian choice*. Springer, 2001.

[Row00] Sam T. Roweis. One microphone source separation. In *Proceedings of NIPS*, pages 793–799, 2000.

[SMD01] H. Snoussi and A. Mohammad-Djafari. Penalized maximum likelihood for multivariate gaussian mixture. In *Proceedings of the Bayesian Inference and Maximum Entropy Methods MaxEnt Workshops*, August 2001.

[STS99] D.W.E. Schobben, K. Torkkola, and P. Smaragdis. Evaluation of blind signal separation methods. In *Proceedings of ICA '99, Aussois, France*, pages 261–266, 1999.

[VM90] A. Varga and R.K. Moore. Hidden markov model decomposition of speech and noise. In *Proceedings of ICASSP*, pages 845–848, 1990.

[Wie49] N. Wiener. *Extrapolation, interpolation and smoothing of stationary time series*. MIT press, 1949.

[WN97] E. Wan and A. Nelson. Neural dual extended kalman filtering : Applications in speech enhancement and monaural blind signal separation. In *Proceedings of the IEEE Workshop on Neural Networks and Signal Processing*, 1997.

[WW93] Eva Wesfreid and Mladen Victor Wickerhauser. Adapted local trigonometric transform and speech processing. *IEEE Transactions on Signal Processing*, 41(12) :3596–3600, 1993.

Résumé :

Le problème de la séparation de sources sonores dans des conditions quasi-réelles, c'est-à-dire avec peu de microphones, suscite un intérêt croissant de la part de la communauté du traitement du signal. Dans ce cadre, le problème de séparation de sources avec un capteur unique a été peu étudié et est en émergence.

L'objet de cette thèse est l'étude de la séparation de sources sonores avec un seul capteur, dans le domaine temps-fréquence. Deux méthodes permettant de résoudre partiellement ce problème sont exposées d'un point de vue théorique et illustrées sur des exemples réels. Ces méthodes permettent d'étendre les techniques classiques de filtrage fréquentiel (filtrage de Wiener) à des signaux non-stationnaires et en particulier dans le cadre de signaux stationnaires à court terme. La première méthode est une extension du filtrage de Wiener à des modèles de mélange de gaussiennes pour les sources. La seconde méthode est fondée sur une décomposition non négative du spectre du mélange sur un dictionnaire de formes spectrales caractéristiques des deux sources.

Outre des contributions au niveau de la formalisation et de l'algorithmique, un des apports notables des travaux présentés dans cette thèse concerne l'utilisation de modèles à facteur d'amplitude et leur application pour la modélisation de sources stationnaires à court terme.

D'autre part, une évaluation comparative des algorithmes proposés dans cette thèse est fournie sur un ensemble de signaux sonores réels et les modalités d'implémentation des algorithmes sont traités.

Cette thèse est donc une exploration des possibilités d'extensions du filtrage de Wiener pour la séparation de sources avec un seul capteur et c'est aussi, nous l'espérons, un point de départ pour de nouveaux développements tant théoriques que pratiques.

Mots Clefs :

- statistiques bayésiennes - filtrage de Wiener - traitement du signal sonore - séparation de sources - représentations redondantes et décompositions parcimonieuses