Séparation de sources pour l'audition des robots

Mounira Maazaoui

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2012-ENST-016

EDITE ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

Télécom ParisTech

Spécialité “ Signal et Images ”

présentée et soutenue publiquement par

Mounira MAAZAOUI

le 4 mai 2012

Séparation de sources pour l’audition des robots

Directeur de thèse : **Yves GRENIER**
Co-encadrement de la thèse : **Karim ABED-MERAIM**

Jury

M. Rémi GRIBONVAL, Directeur de Recherche, INRIA Rennes	Rapporteur
M. Christian JUTTEN, Professeur, Université Joseph Fourier	Rapporteur
M. Yannick DEVILLE, Professeur, Université Paul Sabatier	Président du Jury
M. Olivier WARUSFEL, Docteur, IRCAM	Examinateur
M. Karim ABED-MERAÏM, Maître de conférences, Télécom ParisTech	Directeur de thèse
M. Yves GRENIER, Professeur, Télécom ParisTech	Directeur de thèse

T H È S E

Télécom ParisTech
Grande école de l’Institut Télécom – membre fondateur de ParisTech
46, rue Barrault – 75634 Paris Cedex 13 – Tél. + 33 (0)1 45 81 77 77 – www.telecom-paristech.fr

Remerciements

Mes plus vifs remerciements s'adressent à mes directeurs de thèse Yves Grenier et Karim Abed-Meraïm. Outre leurs compétences scientifiques solides et variées, ils ont fait preuve d'une excellente pédagogie ainsi que d'une écoute et une disponibilité constantes. Je remercie Yves et Karim surtout pour la confiance qu'ils m'ont accordée tout au long de cette thèse, confiance qui s'est traduite par une liberté dans mon travail. Je les remercie aussi pour leur patience et leur soutien infaillibles.

Je tiens à remercier tout particulièrement les membres du jury pour l'intérêt qu'ils ont porté à mes travaux et pour m'avoir fait l'honneur de participer à ma soutenance de thèse : Rémi Gribonval et Christian Jutten, rapporteurs, Yannick Deville, président du jury et Olivier Warusfel, examinateur.

Je tiens aussi à exprimer mon infinie reconnaissance à Slim Essid pour son soutien constant et pour m'avoir donné l'opportunité d'intégrer Télécom ParisTech et l'excellente équipe Audiosig, je lui dois beaucoup. Je dois également beaucoup à Meriem Jaidane et Sonia Djaziri Larbi. Leur excellente pédagogie a su me mettre sur la voie du traitement de signal audio.

Je remercie tous nos partenaires du projet Romeo, en particulier Aldebaran Robotics et les partenaires du groupe système auditif et vocal.

J'ai passé ces années de thèse dans un environnement exceptionnel grâce à mes amis de TSI, grâce à leur bonne humeur et leur soutien. Je remercie chaleureusement mes amis et collègues de l'équipe Audiosig : Bertrand David (merci pour ton soutien et ton amitié !), Gaël Richard, Roland Badeau, Antoine Liutkus, Rémi Foucard, Manuel Moussallam, Benoit Fuentes, Sébastien Fenet, François Rigaud, Nicolas Lopez, Thomas Fillon, Aymeric Masurelle, Angélique Dremeau, Gaël Ladreyt, Laure Cornu, Honoré Takeugming, Anne-Claire Conneau et Cécilia Damon ; sans oublier les anciens membres de l'équipe audio : Romain Hennequin, Cédric Fevotte, Alexey Ozerov, Cyril Joder, Félicien Vallet et Benoit Mathieu. Des amitiés se sont aussi crées à l'extérieur de l'équipe Audiosig, je pense en particulier à Jonathan Sillan et Romain Bouqueau (merci pour tous ces rires !), Harlina Daud, Manel Abid, Bertha

Rodriguez, Marc Décombas, Jean-Claude Dufourd, Cyril Concolato, Jean Lefeuvre, Jérémy Lardon et Fabrice Planche.

Je remercie en particulier Alya Mlaïki et Aurélien Van Roy d'avoir fait le déplacement de loin afin d'assister à ma soutenance de thèse, Krystian qui m'a beaucoup soutenue durant cette épreuve ainsi que tous mes amis qui ont toujours été présents pour moi.

Enfin, je ne remercierai jamais assez ma famille, en particulier ma soeur Mouna, ma grand-mère et surtout mes chers parents, Taoufik et Faouzia, de m'avoir toujours soutenue, encouragée et mise dans des conditions idéales pour réussir.

Résumé

Le principal objectif de cette thèse est de proposer des algorithmes de séparation aveugle de sources audio en utilisant un réseau de capteurs (plus que deux capteurs). L'application finale de ces algorithmes est la séparation de sources audio pour l'audition des robots dans le cadre du projet Romeo.

Dans cette thèse, nous avons dans un premier temps développé et étudié des algorithmes de séparation aveugle de sources audio basés sur des critères de parcimonie. Nous nous sommes intéressés à la minimisation de la norme l_1 avec une technique d'optimisation du gradient naturel. Cette méthode de séparation a des résultats comparables à ceux obtenus avec l'analyse en composantes indépendantes (ICA) utilisant la même technique d'optimisation. Cette étude nous a menés à nous intéresser de plus près aux critères de parcimonie et nous avons développé un critère basé sur la paramétrisation de la pseudo-norme l_p, avec $0 < p < 1$: ceci revient à rendre la contrainte de parcimonie plus dure au fur et à mesure que l'algorithme avance dans ses itérations. Cette méthode montre des résultats assez prometteurs.

Ensuite, nous avons exploité l'aspect multicapteurs de notre application (16 capteurs sont fixés autour de la tête de l'humanoïde) et nous avons proposé un algorithme de séparation avec une étape de prétraitement de formation de voies fixe. La formation de voies permet de réduire l'effet de la réverbération et du bruit sur les signaux et préparer à la séparation. Les filtres de formations de voies sont calculés hors ligne. Les signaux filtrés par formation de voies deviennent ensuite les entrées d'un algorithme classique de séparation aveugle de sources, dans notre cas, nous avons utilisé celui se basant sur la minimisation de la norme l_1. Dans le cas de l'audition des robots, les capteurs sont souvent placés sur la tête de l'humanoïde. Afin de tenir compte de l'influence de la tête sur le champ sonore proche, nous avons construit les filtres de formation de voies en utilisant les fonctions de transfert de tête (HRTF) du robot. L'étape de formation de voies améliore les résultats de séparation par rapport à l'utilisation d'un algorithme de séparation seule.

Cette thèse propose aussi les versions itérative et adaptative de ces algorithmes. Dans la partie adaptative, nous nous plaçons dans le scénario réel où le nombre de sources présentes dans l'environnement du robot est inconnu et change au cours du temps.

Dans cette thèse, nous avons aussi développé une base de données de HRTF pour l'estimation des filtres de formation de voies et une base de réponses impulsionnelles enregistrées dans deux milieux différents afin d'évaluer les algorithmes de séparation de sources proposés et les comparer à ceux de l'état de l'art. L'évaluation des

performances de ces algorithmes a été faite en utilisant une boite à outils de mesures objectives de la qualité des sources estimées (BSS-EVAL) et une boite à outils de mesures perceptuelles (PEASS).

Abstract

The main objective of this thesis is to propose blind audio source separation algorithms using a microphone array (more than two sensors). The final application of these algorithms is audio source separation for robot audition through the Romeo project.

In this thesis, we first developed and studied blind source separation algorithms based on a sparsity criterion. Our interest was focused on l_1 minimization using the natural gradient optimization technique. The separation performance of this method is close to that obtained by independent component analysis using the same optimization technique. This study led us to be more interested in the sparsity criteria, and we developed an algorithm based on the parametrization of the quazi-norm l_p, with $0 < p < 1$: the sparsity criterion gets harder through the iterations of the algorithm. This separation method shows promising results.

Then, we exploited the multisensor aspect of our application (16 sensors are fixed on the head of the humanoid) and we proposed a separation algorithm with a fixed beamforming preprocessing step. Beamforming reduces the reverberation and noise effect and prepares the separation. The beamforming filters are estimated off-line. After the beamforming filtering, the signals become the input of a classical blind source separation algorithm, in our case we used the one based on the l_1 norm minimization. In the robot audition case, the sensors are often placed on the head of the humanoid. To take into account the influence of the head in the near sound manifold, we built the beamforming filters using the head related transfer functions (HRTF) of the robot. The beamforming step improves the separation results compared to the use of a blind source separation only.

This thesis also proposes the iterative and adaptive versions of those algorithms. In the adaptive part, we consider the real scenario where the number of sources is unknown and changes.

In this thesis, we also developed a HRTF database for the estimation of the beamforming filters and a database of impulse responses recorded in two different rooms for the evaluation of the proposed separation algorithms. The evaluation of the performance of those algorithms was done using objective and perceptual measures of the quality of the separated sources using respectively the BSS-EVAL toolbox and PEASS toolbox.

Table des matières

Remerciements **1**

Résumé **3**

Abstract **5**

Table des matières **7**

Liste des symboles **11**

I Introduction et préalable 15

1 Introduction générale **17**

1.1 Contexte général : Projet ROMEO/Audition des robots 17

1.1.1 Analyse de scènes auditives 18

1.1.2 Analyse computationnelle de scènes auditives 18

1.2 Problématique : Séparation aveugle de sources audio 19

1.3 Objectifs 21

1.3.1 Objectif du projet ROMEO 21

1.3.2 Objectif de cette thèse 23

1.4 Contributions 23

1.4.1 Bases de données pour la séparation de sources 23

1.4.2 Algorithmes de séparation de sources 25

1.5 Organisation du document 27

2 Etat de l'art de la séparation aveugle de sources audio **29**

2.1 Formulation du problème 29

2.1.1 Modèle des signaux 29

2.1.2 Les problèmes relatifs à la séparation de sources dans le domaine fréquentiel . . . 31
2.2 Séparation aveugle de sources audio . . . 32
2.2.1 Algorithmes basés sur l'indépendance des sources . . . 33
2.2.2 Algorithmes basés sur la non-corrélation des sources . . . 34
2.2.3 Algorithmes basés sur la parcimonie dans le domaine temps-fréquence . . . 35
2.2.4 Algorithmes basés sur l'analyse de scènes sonores et la psychoacoustique . . . 36
2.3 Séparation de sources pour l'audition des robots . . . 36
2.3.1 Les premiers essais . . . 36
2.3.2 Utilisation des différences intéraurales d'intensité et de phase . 37
2.3.3 Séparation de sources à deux étapes . . . 38
2.3.4 Localisation et séparation . . . 38
2.3.5 Le système d'audition complet HARK . . . 39

II Séparation de sources basée sur l'information spatiale et structurelle des signaux 41

3 Formation de voies 43
3.1 Formation de voies : principe . . . 44
3.2 Formation de voies adaptative . . . 47
3.2.1 Capon ou MVDR . . . 47
3.2.2 Maximisation du rapport signal sur bruit . . . 47
3.3 Formation de voies fixe . . . 48
3.4 Les fonctions de transfert de tête (HRTF) . . . 49
3.5 Formation de voies fixe en utilisant les HRTF . . . 51
3.5.1 Vers la modélisation de la variété du réseau de capteurs . . . 51
3.5.2 Estimation des filtres de formation de voies par les HRTF . . 53

4 Séparation basée sur l'information structurelle des sources 55
4.1 L'algorithme d'optimisation du gradient naturel . . . 56
4.2 Analyse en composantes indépendantes . . . 57
4.3 Minimisation de la norme l_1 . . . 60
4.4 Minimisation de la pseudo-norme l_p paramétrée . . . 62
4.4.1 Principe . . . 62

4.4.2 Algorithme proposé 63

III Séparation de sources à deux étapes : combinaison de formation de voies et d'algorithme de séparation de sources 67

5 Séparation avec prétraitement par formation de voies 69

5.1 Séparation de sources à deux étapes : principe 70
5.2 Prétraitement avec une formation de voies fixe 72
5.2.1 Formation de voies vers les directions d'arrivées 72
5.2.2 Formation de voies vers des directions de visée fixes 73
5.2.3 Formation de voies vers des directions de visée fixes avec sélection de lobes 74
5.3 Estimation du nombre de sources et des directions d'arrivées 75

6 Séparation adaptative avec une étape de formation de voies 81

6.1 Schéma d'adaptation 82
6.1.1 Fenêtres d'analyse 82
6.1.2 Principe d'adaptation 83
6.1.3 Problèmes de permutation et d'échelle dans le domaine temporel 85
6.2 Algorithme de séparation sans estimation du nombre de sources . . . 85
6.3 Algorithme de séparation avec estimation du nombre de sources . . . 86
6.3.1 Activation d'une ou plusieurs sources 88
6.3.2 Extinction d'une ou plusieurs sources 88

IV Expériences et résultats 91

7 Bases de données pour la séparation de sources 93

7.1 Le réseau de capteurs pour les mesures 93
7.2 Réponse impulsionnelle acoustique et temps de réverbération 94
7.3 Construction des mélanges convolutifs 96
7.4 Calcul des réponses impulsionnelles avec les séquences complémentaires de Golay 98
7.5 Base de données des réponses impulsionnelles dans un milieu réverbérant 100
7.5.1 A Télécom ParisTech 100
7.5.2 A l'institut de la vision 100

7.6 Base de données de HRTF 101
7.6.1 Description matérielle et logicielle 102
7.6.2 Processus expérimental 103

8 Outils d'évaluation des algorithmes de séparation de sources 105
8.1 Évaluation objective des performances de séparation (BSS_EVAL) . 106
8.1.1 Décomposition de la source estimée 106
8.1.2 Mesures des performances globales 108
8.2 Évaluation perceptuelle des performances de séparation (PEASS) . . 108
8.2.1 Modélisation et estimation des composantes de distorsions . . 109
8.2.2 Mesures objectives 111

9 Évaluation des algorithmes itératifs 113
9.1 Comparaison entre le critère de parcimonie et le critère d'indépendance 114
9.2 Minimisation de la norme l_p paramétrée 116
9.3 Évaluation de l'apport du prétraitement par formation de voies des algorithmes de séparation à deux étapes 121
9.3.1 Influence du prétraitement par formation de voies 121
9.3.2 Influence de la séparation angulaire entre les lobes : 126
9.3.3 Influence de la sélection de lobes 130
9.3.4 Analyse de la convergence 136
9.4 Variation des performances de séparation avec le nombre de capteurs 137

10 Évaluation des algorithmes adaptatifs 147
10.1 Evaluation des algorithmes adaptatifs de séparation sans estimation du nombre de sources 148
10.1.1 Nombre de sources connu 148
10.1.2 Nombre de sources fixé *a priori* 152
10.2 Evaluation des algorithmes de séparation avec estimation du nombre de sources 156
10.3 Résumé des résultats 161
10.4 Comparaison avec HARK 165
10.4.1 HARK : principe 165
10.4.2 Evaluation des résultats 167

11 Conclusion et perspectives 171

Bibliographie 175

Table des figures **187**

Index **196**

Liste des symboles

APS Score perceptuel relatif aux artéfacts (Artifacts-related Perceptual Score)

CASA Analyse computationnelle de scènes auditives (Computational Auditory Scene Analysis)

DOA Directions d'arrivées (Directions of Arrival)

EDC Courbe de décroissance de l'énergie (Energy Decay Curve)

HRIR Réponse impulsionnelle de tête (Head Related Impulse Response)

HRTF Fonction de transfert de tête (Head Related Transfer Function)

ICA Analyse en composantes indépendantes (Independent Component Analysis)

IID Différence d'intensité interaurale (Interaural Intensity Difference)

IPS Score perceptuel relatif aux interférences (Interference-related Perceptual Score)

ITD Différence de temps interaurale (Interaural Time Difference)

MVDR Minimum Variance Distortionless Response

OPS Score perceptuel global (Overall Perceptual Score)

SAR Rapport sources-à-artéfacts (Sources-to-Artifacts Ratio)

SDR Rapport signal-à-distorsion (Signal-to-distorsion Ratio)

SIR Rapport source-à-interférences (Sources-to-Interferences Ratio)

SNR Rapport sources-sur-bruit (Sources-to-Noise Ratio)

TDOA Différence de temps d'arrivée (Time Difference Of Arrival)

TFCT Transformée de Fourier à Court Terme

TFCTI Transformée de Fourier à Court Terme Inverse

TPS Score perceptuel relatif à la cible (Target-related Perceptual Score)

Première partie

Introduction et préalable

Chapitre 1

Introduction générale

1.1 Contexte général : Projet ROMEO/Audition des robots

M. Robert, un retraité de 70 ans, est assis sur son fauteuil dans son appartement parisien en écoutant la radio. Par cette chaude matinée du mois de juillet, M. Robert a soif. Mais depuis qu'il est en perte d'autonomie, de simples tâches comme aller chercher un verre d'eau sont de véritables défis pour lui. Mais plus maintenant. "Romeo ! Apporte-moi un verre d'eau". Un robot humanoïde, Romeo, se déplace du séjour vers la cuisine et lui apporte un verre d'eau. Ceci est un des scénarios du projet ROMEO [7] qui constitue le cadre général de cette thèse. Le projet ROMEO vise à développer un robot humanoïde destiné à l'aide aux personnes âgées, malvoyantes ou en perte d'autonomie dans leur vie quotidienne. Le projet ROMEO est labellisé par le pôle de compétitivité Cap Digital et financé par la région Ile-de-France, la Direction Générale de la Compétitivité, de l'Industrie et des Services (DGCIS) et de la ville de Paris.

Le robot du nom de Romeo doit aider son "maître" au quotidien tout au long de la journée dans différentes tâches comme ouvrir la porte d'entrée, lui apporter des objets ou encore le secourir en cas de chute. L'interaction entre Romeo et l'Homme se fait via la voix qui représente une interface facile et accessible au plus grand nombre d'utilisateurs. L'exécution de l'ordre du maître par le robot se base essentiellement sur l'*écoute* et la *compréhension* de cet ordre qui traduisent un comportement proche de celui de l'être humain.

1.1.1 Analyse de scènes auditives

Un humain avec une audition saine est capable de différencier les sons qui arrivent mélangés à ses oreilles et peut se concentrer sur un son en particulier dans un environnement bruyant, l'identifier et le comprendre : c'est l'effet *cocktail party*. Pour reconnaître les composantes du son qui forment le mélange audio arrivant à nos oreilles, le système auditif doit en quelque sorte créer des descriptions basées seulement sur ces composantes qui ont pour origine le même évènement sonore. Le processus qui permet de réaliser cette tâche s'appelle *analyse de scène auditive*.

Le terme "analyse de scènes" a été utilisé pour la première fois par des chercheurs en vision par ordinateur. Il fait référence à la stratégie avec laquelle un ordinateur tente de mettre ensemble toutes les propriétés visibles (contours, textures des surfaces, couleurs, etc...) qui appartiennent au même objet, dans une photographie d'une scène où les parties visibles de cet objet sont discontinues (à cause d'un obstacle se trouvant entre la caméra et l'objet en question). Et ce n'est qu'après ce rassemblement que la forme et les propriétés globales de cet objet sont déterminées. Par analogie selon Bregman [17], l'analyse de scènes auditives est le processus par lequel le système auditif d'un être humain organise le son en des éléments perceptuels significatifs, puis les fusionne ou les sépare afin de distinguer entre les sources présentes dans son environnement. Le concept d'analyse de scènes auditives a été introduit pour la première fois par Bregman en 1990 [17].

1.1.2 Analyse computationnelle de scènes auditives

Dans le scénario présenté au début de cette section, l'humanoïde Romeo est équipé de microphones par analogie aux oreilles humaines. Les microphones de Romeo reçoivent deux signaux audio se trouvant dans l'environnement du robot : la voix du maître et le signal de la radio arrivent aux capteurs mélangés. Un être humain se serait naturellement concentré sur la voix du maître, grâce aux mécanismes de psychoacoustique que nous venons de citer [17]. Pour qu'il puisse agir en conséquence des évènements qui se produisent, le robot doit comprendre son environnement sonore, séparer et localiser les sources, identifier le locuteur, comprendre ce qu'il lui dit et détecter ses émotions : c'est la définition de *l'audition des robots*. L'audition des robots se base sur la modélisation informatique de l'analyse de scènes auditives connue sous le nom d'*analyse computationnelle de scènes auditives* (CASA : Computational Auditory Scene Analysis). L'analyse computationnelle de scènes auditives représente un cadre général du traitement des signaux audio qui vise à comprendre

un mélange arbitraire de sons contenant différents types de signaux (de la parole, des signaux autres que de la parole, des signaux musicaux, etc.) dans des environnements acoustiques différents. Un algorithme de CASA analyse les mélanges audio et doit être capable de dire quelle partie de ce mélange est pertinente pour des problèmes comme la segmentation de flux, l'identification et la localisation des sources mais aussi, et c'est la partie qui nous intéresse dans cette thèse, la séparation des sources.

1.2 Problématique : Séparation aveugle de sources audio

Dans le scénario pilote présenté dans la section précédente, M. Robert donne un ordre à Romeo tout en écoutant la radio. La tâche effectuée par l'humanoïde Romeo peut être décomposée en sous-tâches :

1. Romeo écoute la phrase prononcée par M. Robert.
2. Romeo comprend l'ordre de son maître.
3. Romeo exécute l'ordre de son maître.

La voix de M. Robert arrive au robot mélangée avec le signal émis par la radio : pour que Romeo puisse comprendre et exécuter l'ordre donné par son maître, il faut procéder à une séparation de ces signaux.

Notre tâche dans ce projet se focalise sur la séparation aveugle de sources audio par un réseau de microphones (*cf.* figure 1.1). La séparation de sources consiste à estimer les signaux sources à partir de leurs mélanges reçus aux capteurs. Dans le scénario pilote, les conditions dans lesquelles évolue le robot ne sont pas connues : on ne connaît pas le nombre et les positions des sources, le bruit ambiant, le taux de réverbération de la pièce et encore moins les caractéristiques acoustiques des différents chemins sources-microphones. Le système de mélange n'est donc pas connu *a priori*, dans ce cas la séparation est dite *aveugle*.

L'application fixée par le projet ROMEO, l'audition des robots, ainsi que les différents scénarios du projet considèrent l'évolution du robot dans un milieu réel : un appartement ou une maison. Le robot évoluera donc dans un environnement réverbérant. Les mélanges à la sortie des capteurs sont par conséquent des mélanges convolutifs, par opposition aux mélanges instantanés observés dans des environnements dit anéchoïques, sans réverbération, comme les chambres anéchoïques (les chambres sourdes).

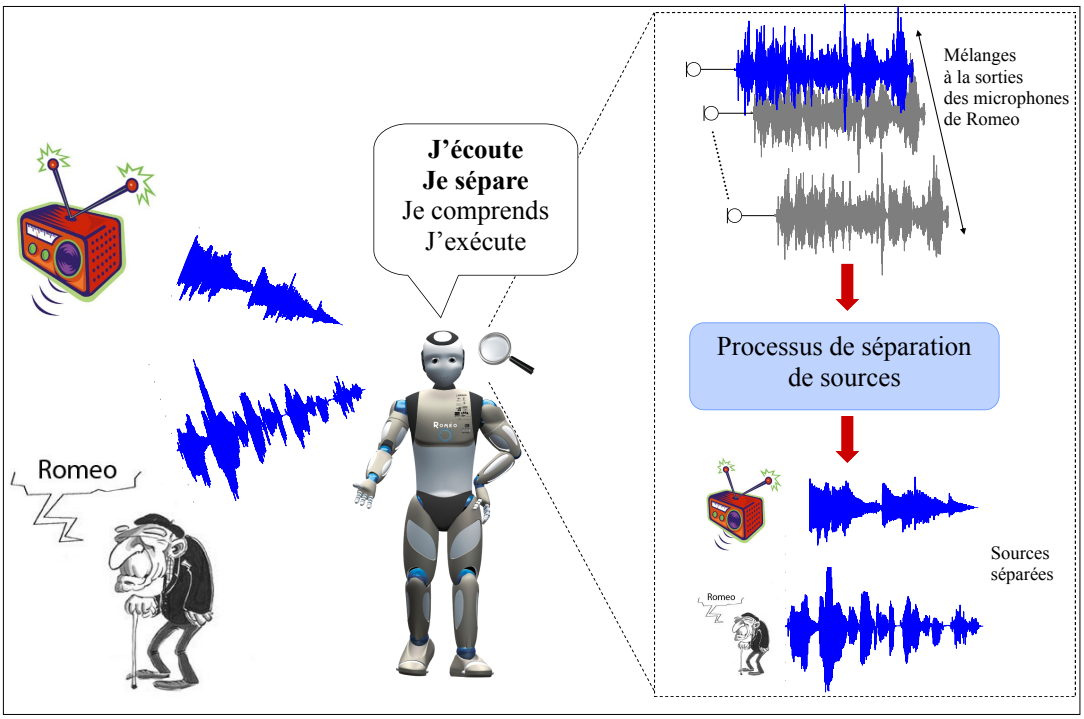


FIGURE 1.1 – Analyse de scènes auditives par Romeo : étape de la séparation de sources

Nous nous plaçons dans un cadre de séparation de sources par un réseau de microphones, avec plus de deux capteurs. En comparant le nombre de sources au nombre de capteurs, la séparation de sources peut être classée en trois cas :

- cas sous-déterminé : nombre de sources supérieur au nombre de capteurs,
- cas déterminé : nombre de sources égale au nombre de capteurs,
- cas sur-déterminé : nombre de sources inférieur au nombre de capteurs.

Dans cette thèse, nous nous intéressons à la séparation de sources sur-déterminée : nous utilisons 16 capteurs et nous supposons que le nombre de sources maximal dans l'environnement du robot est inférieur ou égal à 16.

Plus de détails sur la séparation aveugle de sources audio ainsi qu'un état de l'art des algorithmes de séparation de mélanges convolutifs et ceux relatifs à l'audition des robots seront présentés au chapitre 2.

1.3 Objectifs

1.3.1 Objectif du projet ROMEO

L'objectif du projet ROMEO est de construire un robot humanoïde capable d'aider les personnes en perte d'autonomie en utilisant exclusivement des commandes vocales. Nous nous focalisons sur les objectifs du module audio de ce projet. Ce module comporte quatre parties (*cf.* figure 1.2) :

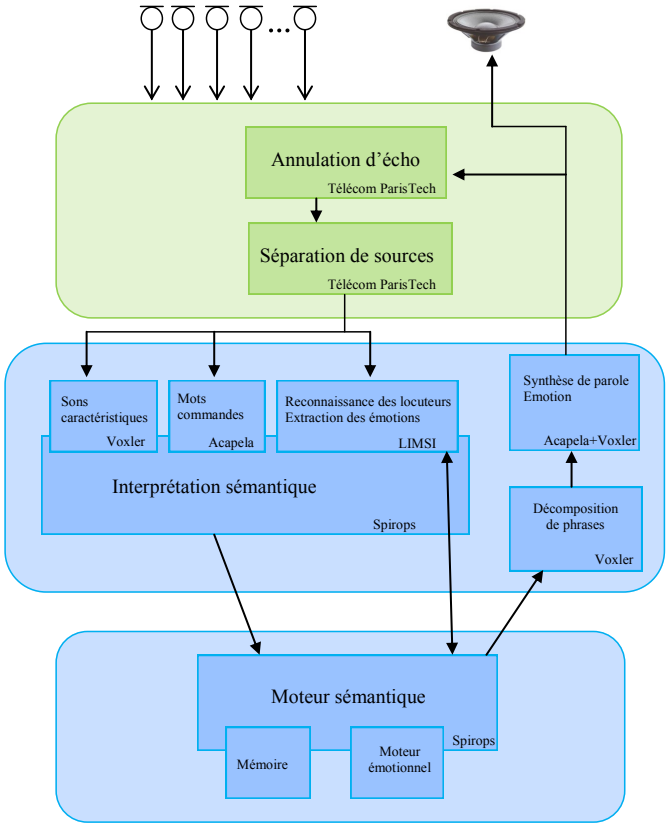


FIGURE 1.2 – Le module de traitement audio du projet ROMEO

Acquisition/Restitution : l'acquisition se fait avec 16 capteurs fixés autour de la tête du robot. Deux des seize capteurs sont équipés chacun d'un pavillon et sont placés à l'intérieur des canaux de ces pavillons pour modéliser les oreilles humaines.

AEC/Séparation/Localisation : c'est la partie la plus importante du module audio et sur laquelle se basent tous les traitements audio à suivre comme la reconnaissance de la parole, des émotions, etc, ... Dans cette partie, nous

effectuons de la localisation et de la séparation de sources. C'est la partie dans laquelle s'inscrit cette thèse, elle sera détaillée dans la section suivante. Notre module de séparation de sources doit s'intégrer au module d'annulation d'écho acoustique (AEC : Acoustic Echo Cancellation) comme le montre la figure 1.3.

Interprétation/Synthèse : l'interprétation consiste en la reconnaissance des locuteurs et des émotions, l'extraction des sons et des bruits caractéristiques (la musique, la sonnette de la porte, etc...), l'extraction d'une transcription écrite de ce que disent les locuteurs et l'extraction de la sémantique de cette transcription. La synthèse consiste en la synthèse de parole et des émotions en réaction à la décision après l'interprétation et la compréhension du contexte faite par le module "Décision".

Décision : à partir de la sémantique extraite dans l'étape "Interprétation" du module "Interprétation/Synthèse", cette partie fournit une décision qui déclenche des comportements.

Nous intervenons dans le module audio du projet ROMEO comme le premier niveau de traitement audio qui consiste en la séparation de sources audio se trouvant dans l'environnement du robot. Notre objectif dans le cadre de ce projet est de fournir un algorithme de séparation aveugle de sources audio capable de traiter les données en temps-réel et de s'adapter au changement dynamique des conditions acoustiques et plus généralement de l'environnement du robot.

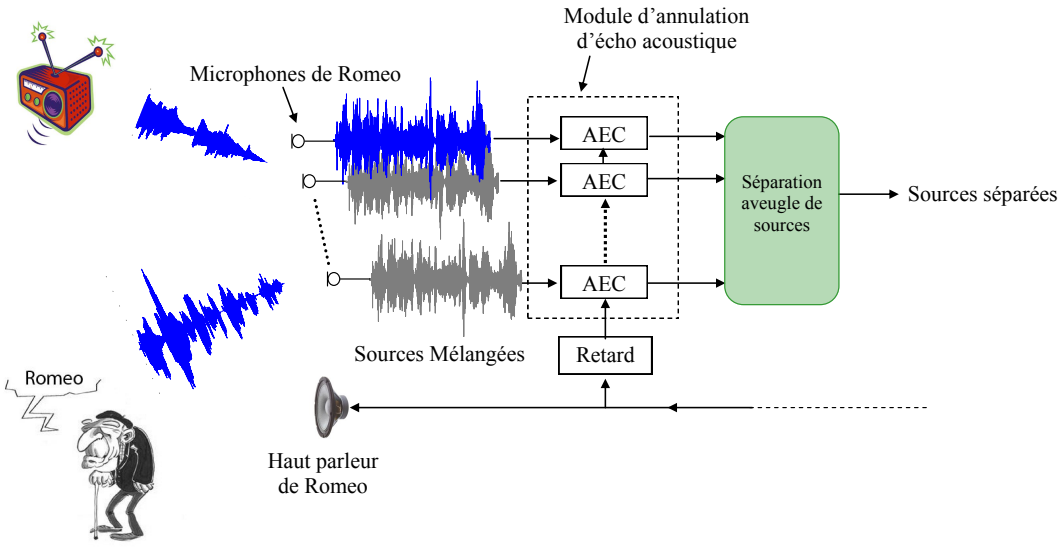


FIGURE 1.3 – Le schéma d'annulation d'écho acoustique

1.3.2 Objectif de cette thèse

L'objectif de cette thèse est de développer de nouveaux algorithmes de séparation aveugle de sources audio dans le contexte de l'audition des robots. Le but est de trouver, en aveugle, des filtres de séparation qui conduisent à l'estimation des sources présentes dans les mélanges reçus aux microphones. Nous proposons de traiter le problème de séparation de sources dans un milieu réel et pour un scénario réel. Nous supposons que nous sommes dans un milieu réel, ceci signifie que les signaux sources arrivent aux capteurs mélangés avec leurs réverbérations. Le mélange est d'autant plus complexe que la pièce dans laquelle évolue le robot est réverbérante. Un scénario réel impose que nous n'ayons aucun *a priori* sur les sources et les conditions du mélange.

Dans cette thèse, nous considérons les scénarios suivant :

- Scénario 1 : le nombre de sources et leurs positions sont fixes au cours du temps,
- Scénario 2 : le nombre de sources changent au cours du temps mais leurs positions restent fixes.

Pour ces deux scénarios, les algorithmes proposés sont évalués en mode itératif (off line) et en mode adaptatif (on line) sous différentes configurations.

1.4 Contributions

1.4.1 Bases de données pour la séparation de sources

Au cours de cette thèse, nous avons développé deux bases de données pour deux applications différentes que nous détaillerons dans les paragraphes suivants. Chacune de ces bases de données a été acquise par 16 capteurs placés autour de la tête d'un mannequin de vitrine de taille enfant mesurant 1m20.

Le prototype de la tête et torse de Romeo prévu pour nos mesures n'a été prêt qu'au mois de novembre 2011 et les premiers tests ne se sont pas révélés concluant pour effectuer l'acquisition des bases de données avec ce réseau de capteurs, ce mannequin de vitrine a été nécessaire pour évaluer les algorithmes de séparation de sources sur une base de données mesurée avec un réseau de capteurs qui modélise celui du robot.

Base de données de fonctions de transfert de têtes (HRTF)

La fonction de transfert de tête (HRTF : Head Related Transfer Function) est une réponse qui caractérise comment un signal source émis d'une direction spécifique est reçu à une oreille. La HRTF de chaque oreille capture l'information de localisation d'une source et la modification introduite par la tête et le pavillon auriculaire sur le chemin de propagation de celle-ci. Les HRTF sont des indices importants pour la perception des sons environnant et la localisation des sources, ils forment le cœur des techniques de spatialisation binaurale. Nous avons généralisé le concept des HRTF au cas d'un robot humanoïde avec plus que deux "oreilles" (plus que deux microphones fixés dans sa tête) et nous proposons Theo-HRTF une base de données de 504×16 HRTF enregistrée avec 16 microphones depuis 72 angles d'azimut et 7 angles d'élévation. Cette base de données des HRTF est disponible en ligne (`http://www.tsi.telecom-paristech.fr/aao/?p=347`), plus de détails sur son acquisition sont donnés dans le chapitre 7. Ces mesures de fonctions de transfert de tête ont été effectuées pour être exploitées dans une formation de voies fixe, une étape de prétraitement proposée avant le module de séparation de sources (*cf.* chapitre 5).

Base de données de réponses impulsionnelles

Pour évaluer les algorithmes de séparation de sources proposés et les comparer aux algorithmes de l'état de l'art les plus pertinents, nous avons développé deux bases de données de signaux enregistrés par le réseau de microphones de Theo dans deux milieux différents. Dans un premier temps, nous avons mesuré les *réponses impulsionnelles* entre différents points d'émission dans la salle et les microphones du réseau de capteurs, ensuite nous considérons une base de données de signaux *bruts* de parole : c'est de la parole enregistrée dans une condition anéchoïque sans aucune influence du milieu d'enregistrement. Pour un nombre de sources et des points d'émission donnés, le mélange à une sortie d'un capteur est obtenu en faisant la somme des convolutions des signaux bruts avec les réponses impulsionnelles entre les positions des points d'émission et le capteur considéré. L'avantage de cette méthode est que nous pouvons varier autant que l'on veut les mélanges en variant seulement les signaux bruts et sans refaire à chaque fois les mesures. Les mesures des réponses impulsionnelles sont faites une seule fois. Pour un point d'émission donné, nous pouvons obtenir plusieurs observations différentes.

Ces mesures ont été faites dans les milieux suivants :

– le studio d'enregistrement de Télécom ParisTech, nous appelons la base de

données enregistrée dans ce milieu Theo-RI-Studio ;
– l'appartement témoin du projet Romeo à l'Institut de la Vision (IDV), nous appelons la base de données enregistrée dans ce milieu Theo-RI-IDV.

1.4.2 Algorithmes de séparation de sources

Dans le cadre de l'audition des robots, nous avons développé un certain nombre d'algorithmes de séparation de sources dans le domaine temps-fréquence. Dans un premier lieu, les algorithmes proposés sont implémentés en mode itératif : le traitement des signaux se fait hors ligne sur l'intégralité des mélanges à séparer. Ensuite, nous proposons la version adaptative de ces algorithmes avec une évaluation dans des scénarios réels où le nombre de sources change dynamiquement.

Minimisation de la norme l_1

Nous avons commencé par explorer la séparation de sources audio en utilisant un critère de parcimonie. La minimisation de la parcimonie des sources mélangées conduit-elle à leurs séparation ? Pour répondre à cette question, nous avons choisi la mesure de parcimonie la plus connue grâce notamment à sa convexité : la norme l_1. Nous avons donc procédé à la minimisation de la norme l_1 par une méthode de gradient naturel afin d'estimer les matrices de séparation et donc les sources. Cette méthode de séparation a de bonnes performances, comparables à celles de l'analyse en composantes indépendantes utilisant le même algorithme d'optimisation.

Minimisation de la pseudo-norme l_p paramétrée

Que se passe-t-il si, au lieu d'utiliser la norme l_1 comme fonction de coût, nous utilisons une pseudo-norme plus contraignante au niveau de la parcimonie à savoir la pseudo-norme l_p avec $0 < p < 1$? Plus le paramètre p de la pseudo-norme l_p est proche de 0, plus cette mesure de parcimonie est rigide. Nous avons utilisé la pseudo-norme l_p comme fonction de coût et nous avons observé les performances de l'algorithme de séparation pour plusieurs valeurs du paramètre p, p étant toujours strictement entre 0 et 1. Nous avons remarqué que le résultat de la séparation dépend de ce paramètre : nous pouvons obtenir de meilleurs résultats de séparation en utilisant $0 < p < 1$. Mais le paramètre p optimal varie d'un cas de mélange à un autre et il est assez difficile de le fixer. Nous avons donc procédé à une "paramétrisation" de la norme l_p : nous faisons décroître le paramètre p de 1 à 0 au cours des itérations de l'algorithme de gradient, ce qui fait durcir la contrainte de parcimonie au fur et à

mesure que l'on descend vers la solution. Cette méthode de séparation présente des résultats prometteurs comme nous le détaillerons dans le chapitre 8.

Combinaison de la formation de voies fixe et d'algorithmes de séparation de sources

La séparation de sources dans un milieu réel reste un problème difficile principalement à cause de la réverbération. Pour limiter la réverbération et par conséquent essayer d'améliorer les performances de séparation, nous proposons d'utiliser une formation de voie fixe comme prétraitement de l'algorithme de séparation de sources. Nous disposons d'un nombre important de capteurs (16 capteurs) ce qui nous permet d'obtenir des diagrammes de directivité assez précis. Cependant, la construction des filtres de formation de voies nécessite la modélisation de la variété du réseau de capteurs. Dans le cas de l'audition des robots, les capteurs sont souvent fixés autour de la tête du robot. Nous proposons de prendre en compte l'influence de la tête sur le champ acoustique environnant pour la construction des filtres de formation de voies. Pour ceci, nous utilisons les fonctions de transfert de tête (HRTF) comme vecteurs directionnels ce qui permet de tenir compte de l'influence de la tête sur le champ sonore dans la construction des filtres de formation de voies. Plus de détails sur la construction des filtres de formation de voies par les HRTF seront présentés dans le chapitre 3. Nous avons développé plusieurs variantes de l'algorithme de prétraitement avec formation de voies qui montrent des performances bien supérieures à celles obtenues par des algorithmes de séparation seuls.

Etude de l'influence du nombre de capteurs sur la performance de séparation

Nous avons étudié l'effet du nombre de capteurs sur la qualité de la séparation de sources. Nous avons évalué la performance de séparation de l'algorithme de séparation avec la minimisation de la norme l_1 et le même algorithme de séparation mais avec une étape de prétraitement avec formation de voies fixe en variant le nombre de capteurs. Nous avons considéré un nombre de capteurs allant du cas binaural jusqu'au cas multicapteurs de 16 microphones. Nous avons tenté de trouver le nombre optimal de capteurs qui doit être utilisé pour l'audition des robots avec une géométrie du réseau de capteurs donnée et nous montrons que l'utilisation d'un réseau de capteurs augmente significativement les performances de séparation par rapport au cas binaural et que cette augmentation se stabilise à partir d'un certain nombre de capteurs. Nous pensons que ce nombre de capteurs à partir du-

quel nous n'avons plus d'augmentation significative du gain dépend des conditions acoustiques de l'environnement de la séparation de sources, en particulier le taux de réverbération.

1.5 Organisation du document

Ce document est organisé en quatre parties :

- une partie introductive qui finira par le chapitre II où nous présentons un état de l'art des principales méthodes de séparations de sources, nous nous intéresserons en particulier à la séparation de sources pour l'audition des robots ;
- dans la deuxième partie, nous nous intéresserons aux méthodes de séparation de sources basées sur l'information spatiale (méthodes de formation de voies) et structurelle (méthodes basées sur la parcimonie des sources dans le domaine temps-fréquence) des sources dans le domaine temps-fréquence ; nous présentons en particulier la formation de voies, l'analyse en composantes indépendantes et la séparation avec un critère de parcimonie ;
- la troisième partie sera consacrée à l'étude de la combinaison de la formation de voies et d'algorithmes de séparation de sources ; nous étudierons l'effet de la formation de voies comme prétraitement d'un algorithme de séparation de sources et ceci avec différentes configurations ; nous élargirons ensuite ce concept à la séparation adaptative de sources en ajoutant la difficulté d'un nombre de sources variable au cours du temps ;
- nous finirons ce rapport de thèse par la quatrième partie consacrée aux expériences et aux résultats où nous détaillerons le processus expérimental et les différents résultats obtenus lors de cette thèse.

Chapitre 2

Etat de l'art de la séparation aveugle de sources audio

Introduction

Dans ce chapitre, nous présentons un état de l'art des principales méthodes de séparation aveugle de sources. Nous nous concentrerons ensuite sur les méthodes qui ont été appliquées dans le cadre de l'audition des robots. Mais tout d'abord, nous commençons par le modèle de mélange et les principales notations que nous adopterons le long de ce rapport.

2.1 Formulation du problème

2.1.1 Modèle des signaux

Supposons que l'on dispose de N sources audio $\mathbf{s}(t) = [s_1(t), \ldots, s_N(t)]^T$ et d'un réseau de M microphones. Les sorties du réseau de capteurs sont notés $\mathbf{x}(t) = [x_1(t), \ldots, x_M(t)]^T$, t étant l'indice de temps. Nous supposons que nous sommes dans un cas de séparation sur-déterminé $M \geq N$. Comme nous travaillons dans un environnement réel, les mélanges à la sortie des microphones dans le domaine temporel sont modélisés comme la somme des convolutions entre les signaux sources et les réponses impulsionnelles des différents chemins de propagation entre les sources et les capteurs (*cf.* figure 2.1).

On note $\mathbf{h} = [\mathbf{h}(0), \ldots, \mathbf{h}(L-1)]$ les réponses impulsionnelles tronquées à la longueur L, où $\mathbf{h}(l) = [h_{ij}]_{1 \leq i \leq M,\, 1 \leq j \leq N}$, est une matrice de dimension $M \times N$ contenant les $l^{ième}$ coefficients des réponses impulsionnelles des différents

chemins acoustiques entre les N sources et M capteurs. Les mélanges à la sortie des capteurs s'écrivent dans le domaine temporel comme suit :

$$\mathbf{x}(t) = \sum_{l=0}^{L-1} \mathbf{h}(l)\,\mathbf{s}(t-l) + \mathbf{n}(t) \tag{2.1}$$

où $\mathbf{n}(t)$ est un vecteur de bruit stationnaire. Nous considérons un bruit diffus, spatialement décorrélé, dont l'énergie est supposée être négligeable par rapport à celle des sources. Si le bruit est ponctuel, il sera considéré comme une source sonore. Ce scénario correspond à notre processus expérimental et à l'application finale.

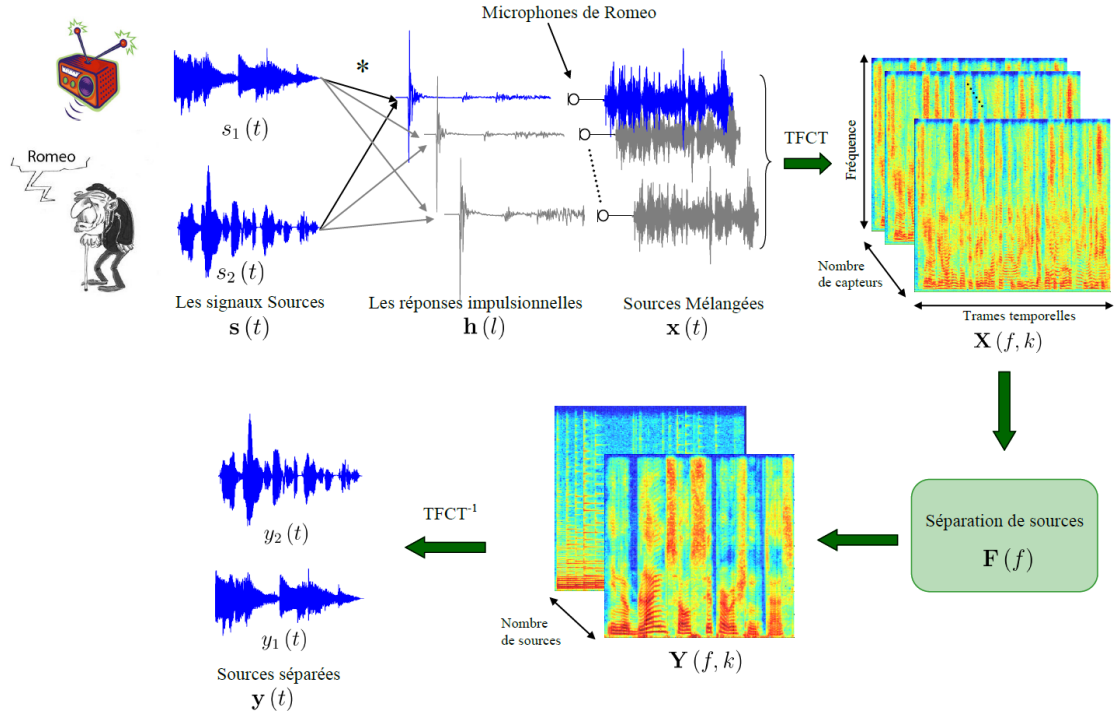


Figure 2.1 – Principe de la séparation de sources

Dans le domaine temporel, le problème de séparation de sources se résume à estimer des filtres de séparation $\mathbf{f} = [\mathbf{f}(0), \ldots, \mathbf{f}(L'-1)]$ de longueur L', avec $\mathbf{f}(l) = [f_{ji}]_{1\leq i\leq N,\, 1\leq j\leq M}$ une matrice de dimension $N \times M$ contenant les $l^{ième}$ coefficients des filtres de séparation des N sources à partir des M capteurs.

Dans le domaine fréquentiel, quand la longueur de la fenêtre d'analyse N_f de la Transformée de Fourier à Court Terme (TFCT) est au moins deux fois plus grande que la longueur L des filtres de mélanges (les réponses impulsionnelles), le mélange convolutif est approché à un mélange instantané à chaque fréquence f. Le problème de séparation devient donc plus simple que celui posé dans le domaine temporel et

nous estimons des matrices de séparation au lieu de filtres de séparation dont la longueur devient importante si nous travaillons avec des signaux réels.

Dans le domaine temps-fréquence, les signaux du mélange à l'indice fréquentiel f et à la trame temporelle k peuvent être approchés par :

$$\mathbf{X}\left(f,k\right) \simeq \mathbf{H}\left(f\right)\mathbf{S}\left(f,k\right) \tag{2.2}$$

$\mathbf{X}(f,k) = [X_1(f,k),\ldots,X_M(f,k)]^H$ est la TFCT de $\{\mathbf{x}\left(t\right)\}_{1\leq t\leq T}$ à la fréquence $f \in \left[1, \frac{N_f}{2}+1\right]$ et la fenêtre d'analyse $k \in [1, N_T]$, $\mathbf{S}(f,k) = [S_1(f,k),\ldots,S_N(f,k)]^H$ est la TFCT de $\{\mathbf{s}\left(t\right)\}_{1\leq t\leq T}$. $\mathbf{H}$ est la transformée de Fourier des filtres de mélanges $\{\mathbf{h}\left(l\right)\}_{0\leq l\leq L-1}$.

Notre objectif est d'utiliser un critère de séparation approprié et de trouver, à chaque fréquence, une matrice de séparation $\mathbf{F}\left(f\right)$ qui conduira à l'estimation des sources originales dans le domaine temps-fréquence :

$$\mathbf{Y}\left(f,k\right) = \mathbf{F}\left(f\right)\mathbf{X}\left(f,k\right) \tag{2.3}$$

où $\mathbf{Y}(f,k) = [Y_1(f,k),\ldots,Y_N(f,k)]^H$. La Transformée de Fourier à Court Terme Inverse (TFCTI) des signaux sources estimés $\mathbf{Y} = \{\mathbf{Y}(f,k)\}_{1\leq f\leq N_f, 1\leq k\leq N_k}$ permet de retrouver les sources $\mathbf{y}\left(t\right) = \left[y_1\left(t\right),\ldots,y_N\left(t\right)\right]^T$ dans le domaine temporel :

$$\mathbf{y}\left(t\right) = \sum_{l=0}^{L'-1} \mathbf{f}\left(l\right)\mathbf{x}\left(t-l\right) \tag{2.4}$$

2.1.2 Les problèmes relatifs à la séparation de sources dans le domaine fréquentiel

La séparation de sources dans le domaine fréquentiel présente un avantage par rapport à la séparation de sources dans le domaine temporel qui est l'estimation de matrices de séparation au lieu de filtres, ce qui engendre des algorithmes de séparation plus simples et donc plus rapides. Cependant, travailler dans le domaine fréquentiel présente deux problèmes bien connus : le problème d'échelle et le problème de permutation. Ces problèmes se traduisent en écrivant l'équation (2.3) des sources estimées dans le domaine temps-fréquence sous la forme :

$$\mathbf{Y}\left(f,k\right) = \mathbf{F}\left(f\right)\mathbf{X}\left(f,k\right) = \mathbf{P}(f)\mathbf{D}(f)\mathbf{S}\left(f,k\right) \tag{2.5}$$

où $\mathbf{P}(f)$ une matrice de permutation et $\mathbf{D}(f)$ est une matrice diagonale d'échelle.

Problème d'échelle A chaque fréquence f, la multiplication des sources estimées par des constantes $\mathbf{D}(f)$ n'affecte le résultat de l'algorithme de séparation que par un filtrage des signaux estimés une fois convertis dans le domaine temporel.

Problème de permutation L'ordre des sources estimées peut être différent à des fréquences adjacentes. Ceci est modélisé par la multiplication des sources originales dans le domaine fréquentiel $\mathbf{S}(f,k)$ par une matrice de permutation $\mathbf{P}(f)$. Si le problème de permutation n'est pas résolu, la reconstruction dans le domaine temporel d'une source donnée contiendra des contributions des autres sources.

2.2 Séparation aveugle de sources audio

Le livre de Pierre Comon and Christian Jutten [24] présente un bon résumé de l'histoire de la séparation de source dont voici les principaux points. Le problème de la séparation de sources a été formulé pour la première fois en 1982 par Bernard Ans, Jeanny Hérault et Christian Jutten dans le cadre de la modélisation neuronale pour le décodage de mouvements dans les vertèbres. A Grenoble en septembre 1987, les acteurs sont J.F. Cardoso, J. Hérault, P. Comon et C. Jutten. Lors d'un workshop, J.F. Cardoso visite J. Hérault et C. Jutten qui lui expliquent le principe de la séparation de sources et lui montrent une démonstration en temps-réel accompli par un appareil purement analogique que C. Jutten a construit en 1985. L'appareil était capable de séparer en temps-réel deux sources audio dans un mélange contrôlé par un potentiomètre. Immédiatement mais indépendamment, J.-F. Cardoso et P. Comon deviennent très enthousiastes à propos de la séparation aveugle de sources audio. Ensuite en 1990, le problème de séparation de sources a été traité par plusieurs chercheurs de plusieurs domaines : traitement de signal, statistique, réseau de neurones, etc, ... En 1991, le premier papier international consacré à ce problème a été publié dans le journal *Signal Processing* et en 1999, la première conférence internationale complètement dédiée à la séparation aveugle de sources a été organisée à Aussois dans les Alpes Français et a attiré 130 chercheurs du monde entier.

Initialement, le problème de séparation de sources a été posé pour les mélanges linéaires instantanés, puis il a été généralisé aux mélanges convolutifs au début des années 90. Puisque nous travaillons avec des signaux enregistrés dans des milieux réels réverbérants, nous nous intéressons aux méthodes de séparation de mélanges convolutifs. Les algorithmes de séparation de mélanges convolutifs peuvent être classés [66] selon :

- le domaine de séparation : domaine temporel, domaine fréquentiel ou domaine temps-fréquence,
- le critère sur lequel se base la séparation : l'indépendance des sources, la non-corrélation des sources, la parcimonie dans le domaine temps-fréquence ou les principes d'analyse de scènes sonores et la psychoacoustique.

Dans cette section, nous allons présenter un aperçu de l'état de l'art de la séparation de sources dans les mélanges convolutifs en les classant selon leurs critères de séparation. Si le lecteur veut plus de détails sur ces méthodes et est curieux d'explorer d'autres méthodes pas nécessairement dans notre sujet, il peut consulter les livres [24, 50].

2.2.1 Algorithmes basés sur l'indépendance des sources

Dans la première classe d'algorithme, la séparation de sources est basée sur l'hypothèse d'indépendance statistique des sources. Nous nous intéresserons dans cette section aux statistiques de $4^{ième}$ ordre et aux méthodes basées sur la théorie de l'information. Nous parlerons aussi de l'Analyse en Composantes Indépendantes (ACI) qui correspond à un cadre général pour résoudre les problèmes de séparation aveugle de sources basés sur l'indépendance statistique des sources à estimer. Elle a été introduite en 1987 par Jutten et formalisée pour les mélanges linéaires par Comon en 1991 [22, 23].

Statistique du 4ème ordre L'indépendance statistique peut se traduire en annulant tous les moments croisés entre les modèles des sources :

$$E\left[y_n(t)^{\alpha}\, y_m(t-\tau)^{\beta}\right] = 0\,,\ n \neq m,\ \{\alpha,\beta\} \subset \mathbb{N},\ \forall\tau \tag{2.6}$$

où $E(.)$ est l'espérance mathématique.

Cependant, pour atteindre la séparation, il n'est pas nécessaire de minimiser tous les moments croisés, plusieurs algorithmes se basent sur la minimisation des statistiques de quatrième ordre. La minimisation des cumulants de quatrième ordre a été étudiée par plusieurs chercheurs [19, 25, 41, 75]. Ye *et al.* ont utilisé les cumulants du second et troisième ordre pour la séparation de sources [96]. L'algorithme le plus connu basé sur cette méthode est JADE de Cardoso et Souloumiac [21] pour les signaux complexes appliqué dans le domaine temps-fréquence pour séparer les mélanges convolutifs. Une mesure des statistiques de quatrième ordre est le kurtosis. Cette mesure de la non-Gaussianité des signaux a été utilisée dans plusieurs algo-

rithmes [77, 82]. D'autres algorithmes utilisent les statistiques d'ordre supérieur avec des fonctions non linéaires, le développement de Taylor de ces fonctions capture les moments d'ordre supérieur et c'est suffisant pour séparer les mélanges convolutifs. [43].

Du côté de la théorie de l'information Dans la théorie de l'information, l'indépendance entre les sources est exprimée en fonction de leurs densités de probabilité. Augmenter l'indépendance entre les sources $\mathbf{y}$ équivaut à minimiser leur information mutuelle ce qui revient à maximiser l'entropie. Le maximum d'entropie est obtenu lorsque la somme des entropies des variables y_n est égale à l'entropie jointe de $\mathbf{y}$. Un algorithme bien connu qui se base sur cette idée est Infomax de Bell et Sejnowski [42] qui a été amélioré en vitesse de convergence par la méthode du gradient naturel d'Amari [13]. L'algorithme Infomax peut aussi être dérivé en utilisant le maximum de vraisemblance [9] ou la divergence de Kullback-Leibler entre les distributions empiriques et le modèle d'indépendance [20]. L'algorithme de Bell et Sejnowski a été étendu au cas convolutif par Torkkola [84].

2.2.2 Algorithmes basés sur la non-corrélation des sources

Dans cette deuxième classe d'algorithmes, la séparation de sources est basée sur l'hypothèse de la non corrélation des sources plutôt que la condition plus forte de leur indépendance. La décorrélation des sources s'obtient en utilisant les statistiques de second ordre, elle n'implique pas l'indépendance et est insuffisante pour la séparation (le système d'équation est dans ce cas sous-déterminé [49]). Cependant, quand les sources sont non-stationnaires, les statistiques de second ordre sont différentes à chaque trame temporelle. Ceci implique que plus d'équations sont disponibles pour résoudre le problème et que le système d'équations pour estimer la matrice de séparation peut être résolu. Les statistiques variables au cours du temps donnent des informations additionnelles pour la séparation : c'est la décorrélation basée sur la non-stationnarité des sources.

Plusieurs algorithmes se basent sur les statistiques de second ordre avec une condition de non-stationnarité des sources. L'idée de la séparation basée sur la décorrélation des signaux non-stationnaires a été proposée par Weinstein *et al.* en 1993 [92]. Ils ont proposé de minimiser les puissances des sources estimées pendant des périodes de stationnarité du signal différentes. Wu et Principe [93] étendent le principe de la diagonalisation multiple à la séparation des mélanges convolutifs dans le domaine temps-fréquence en se basant sur la méthode proposée par Souloumiac

en 1995 pour les mélanges à bande étroite et qui exploite la structure des valeurs propres de deux matrices de covariance [76]. L'utilisation des statistiques de second ordre sur les mélanges non-stationnaires convolutifs s'est poursuivie avec Pham et Cardoso et Matsuka *et al.* qui se basent sur la non stationnarité des sources appliquée aux mélanges instantanés [51, 68]. Cette méthode a été appliquée dans le domaine temporel [45] et fréquentiel [44]. Une version adaptative a été proposée par Aichner *et al.* en 2003 [10]. Un algorithme de séparation bien connu qui se base sur la minimisation de la corrélation croisée des sources à des instants multiples dans le domaine fréquentiel est celui de Parra et Spence [64]. Les statistiques de second ordre sont capturées par le spectre de puissance à des instants multiples et les matrices de mélange et de séparation sont estimées avec une méthode des moindres carrés. Une approche adaptative a été proposée la même année [65]. En 2004, Wang *et al.* proposent une version de ce même algorithme en incorporant une fonction de pénalité dans la fonction de coût [91]. Ceci élimine les contraintes dans une optique de programmation non-linéaire et améliore les résultats.

Si les sources sont non-blanches, nous avons une corrélation décalée pour de multiples délais temporels. Les statistiques de second ordre sont différentes dans chaque décalage temporel et donc plus d'équations sont disponibles et le système d'équation peut être résolu. C'est la décorrélation en se basant sur la non-blancheur des sources ou sur le décalage-temporel [53]. Aichner *et al.* ont combiné l'approche basée sur la non-stationnarité des sources et l'approche basée sur leur blanchiment pour améliorer les résultats de la décorrélation [11, 18].

D'autres méthodes utilisant la non corrélation des signaux sources à estimer se basent sur d'autres principes d'estimation que les statistiques de second ordre tels que la phase minimale du système de mélange et la stationnarité cyclique des signaux sources [66].

2.2.3 Algorithmes basés sur la parcimonie dans le domaine temps-fréquence

Ce type d'algorithme se base souvent sur le non recouvrement des sources dans le domaine temps-fréquence. Les méthodes les plus utilisées sont le masquage temps-fréquence et le clustering.

Nous pouvons voir ce type d'algorithme du point de vue de leur critère : une fonction de coût qui agit sur la parcimonie des mélanges pour les séparer (norme l_p, l_1, etc). [26] présente un aperçu des méthodes de séparation de sources en se basant

sur la parcimonie.

2.2.4 Algorithmes basés sur l'analyse de scènes sonores et la psychoacoustique

Certains algorithmes de séparation de sources se basent sur les idées issues de l'étude du système auditif et les principes de l'audition humaine : c'est l'analyse de scènes auditives computationnelle que nous avons introduite dans le chapitre précédent. Le terme analyse de scènes auditives introduit par Bregman en 1990 [17] fait référence à la capacité d'un humain à former des représentations perceptuelles des sources présentes dans un mélange acoustique comme l'effet "cocktail party". Selon Bregman, l'analyse de scènes auditives est un processus à deux étapes. Dans la première étape, le mélange acoustique est décomposé en éléments qui décrivent un évènement acoustique signifiant. Dans la deuxième étape, les éléments qui peuvent provenir de la même source sont regroupés pour former une structure perceptuelle.

Parmi les outils perceptuels qui ont été utilisés dans la séparation de sources, nous citons [37, 90] qui ont utilisé le masquage perceptuel pour résoudre le problème de permutation, [83] qui a utilisé le pitch dans un algorithme de séparation semi-aveugle pour séparer les sources pendant les parties voisées, [67] qui a utilisé les fonctions de transfert de tête pour poser des contraintes géométriques à la séparation et [71] qui a utilisé la différence interaurale d'intensité (IID : Interaural Intensity Difference) et la différence interaurale de phase (IPD : Interaural Phase Difference) pour différencier les sources.

2.3 Séparation de sources pour l'audition des robots

Dans cette section, nous présentons l'état de l'art des algorithmes de séparation de sources qui ont été publiés spécifiquement dans le cadre de l'audition des robots.

2.3.1 Les premiers essais

Une première discussion sur l'importance d'un module de traitement de signal audio dans un robot autonome a été lancée en 1994 par Brooks et Stein [8]. Les auteurs proposent une réflexion sur un système physique intégrant la vision, une entrée/sortie audio et des manipulations adroites, non seulement dans un but d'ingénierie pour construire un prototype général d'un robot autonome mais aussi dans un but scientifique pour comprendre la cognition humaine.

En 1997, Huang *et al.* proposent un système de localisation et de séparation de sources en utilisant 3 capteurs installés verticalement en haut d'un robot [39]. La localisation est effectuée grâce à la différence de temps d'arrivée des sources par rapport aux deux capteurs qui reçoivent la plus grande énergie des signaux. Ce robot fait de l'*audition active* : il modifie son comportement par rapport aux conditions d'écoute, par exemple il se dirige vers la source si celle-ci est loin pour mieux la localiser. Quand à la séparation de sources, elle se fait grâce à la disparité de phase. Les auteurs ne sont pas rentrés dans les détails concernant ces techniques. La séparation a été faite sur 2 signaux de parole séparés en azimut de 38° dans un milieu anéchoïque et une salle réverbérante. Les auteurs affirment avoir obtenu un bon résultat de séparation dans les deux milieux en se basant sur des mesures subjectives (l'allure des signaux séparés et l'écoute).

2.3.2 Utilisation des différences intéraurales d'intensité et de phase

Si le robot dispose de deux microphones placés au niveau des oreilles humaines, on peut penser à utiliser la différence interaurale d'intensité (IID : Interaural Intensity Difference) et la différence interaurale de phase (IPD : Interaural Phase Difference) pour résoudre le problème de localisation et de séparation des sources. C'est ce qu'ont fait en 2003, Nakadai *et al.* [57]. Ces auteurs utilisent la localisation par IID et IPD afin d'extraire les sources par un filtre passe direction actif qui sépare le signal émis d'une direction bien déterminée. Pour estimer les IID et IPD tout en n'ayant pas recours au calcul des fonctions de transferts de tête comme l'ont fait Matsusaka *et al.* [52], les auteurs proposent deux méthodes. Tout d'abord, ils ont appliqué à l'audio la géométrie épipolaire connue dans le traitement des signaux vidéos [56]. Cette méthode leur a permis d'estimer l'IPD en continue pour toutes les directions d'arrivées des sources et l'IID pour seulement le centre, la droite et la gauche du robot. Ensuite, les auteurs ont utilisé la théorie de la diffraction : le champ à proximité des capteurs de droite et de gauche est calculé en tenant compte du champ incident et du champ diffracté. Les IID et IPD sont ainsi calculés en tenant compte de l'effet qu'a la tête sur ce champ sonore, c'est à dire en tenant compte de la diffusion. L'évaluation de cette technique a été effectuée sur le robot SIG, avec un son harmonique composé de 30 fréquences de 100 Hz à 3000 Hz dans une chambre dont le temps de réverbération est de $300ms$. Les auteurs affirment que les résultats de localisation et de séparation obtenus en utilisant la théorie de la diffraction sont

comparables à ceux obtenus en utilisant les fonctions de transfert de tête.

2.3.3 Séparation de sources à deux étapes

Valin *et al.* [85] proposent en 2004 un système de séparation de sources à deux étapes pour l'audition des robots. La première étape consiste en une séparation de sources géométrique (GSS : Geometric Source Separation) basée sur l'algorithme de Parra et Alvino [63]. Cet algorithme consiste en la décorrélation des signaux à la sortie des capteurs en imposant une contrainte de gain unitaire dans la direction des sources visées et des zéros dans la direction des sources interférentes. La deuxième étape consiste en un filtrage multi-canal utilisant l'estimation du bruit ambiant et les sources interférentes pour améliorer le signal produit lors de la séparation géométrique [54]. Cet algorithme a été appliqué à Pioneer 2, un robot avec 8 capteurs pour séparer 3 sources. Cette méthode a été utilisée en 2007 comme un prétraitement de la reconnaissance de la parole dans le robot SIG 2 [86].

En 2005, Saruwtari *et al.* proposent aussi un algorithme de séparation à deux étapes [73]. Les signaux binauraux observés aux "oreilles" d'un humanoïde sont traités d'abord avec un modèle à une seule entrée et multiple sorties SIMO (Single-Input and Multiple-Output) basé sur ICA [79] et ensuite avec un masquage binaire [31]. Les expériences ont été faites avec la tête et le torse d'un mannequin dans une salle dont le temps de réverbération est de 200 ms et avec deux sources. Les auteurs montrent que l'utilisation du modèle SIMO basé sur ICA avec masquage binaire donne de meilleurs résultats de séparation que l'utilisation d'une ICA conventionnelle suivie du même masquage.

2.3.4 Localisation et séparation

Tamai *et al.* se sont intéressés à la séparation de sources pour l'audition des robots en utilisant deux configurations de réseau de capteurs différentes. En 2004, ils proposent un réseau de 32 microphones fixés sur un cercle de 50 cm de diamètre et qui peut être monté sur un robot mobile. Dans un premier temps, ils proposent de localiser les sources par une formation de voies type *delay-and-sum* [80]. En 2005, une partie de l'équipe précédente change la configuration du réseau de capteurs qui devient un réseau de 3 anneaux avec le même nombre de microphones [81]. La localisation est effectuée par la formation de voies delay-and-sum et la séparation est faite à la suite par un algorithme de sélection de bande de fréquence [74].

En 2007, Rudzyn *et al.* décrivent un système d'audition des robots nommé RRAS

[70]. Avec un réseau de 4 capteurs, ils procèdent à une localisation 3D et se focalisent sur la source d'intérêt. Cette focalisation se fait suite à une caractérisation de la voix leur permettent de distinguer si la source sonore détectée est de type parole ou pas. La même année, [38] utilise ICA pour la séparation de source pour l'audition des robots avec un module de reconnaissance du locuteur.

2.3.5 Le système d'audition complet HARK

Yamamoto *et al.* proposent une technique de séparation de sources basée sur des contraintes géométriques comme un prétraitement au module de reconnaissance de la parole de leur système d'audition des robots [86, 94, 95]. Ce système a été implémenté dans les humanoïdes SIG2 et ASIMO avec un réseau de 8 microphones, comme un module d'un système d'audition des robots plus complet nommé HARK [59]. La référence [62] propose une analyse de ce problème d'un point de vue analyse computationnelle de scènes auditives en utilisant l'algorithme MUSIC pour la localisation des sources et l'algorithme proposé par Parra et Alvivo [63] pour la séparation.

De 2008 à 2010, Nakajima *et al.* dans [58, 60, 61] présentent un algorithme de séparation de sources géométrique en réglant automatiquement le pas d'avancement et le poids de l'algorithme d'optimisation. Cet algorithme a été implémenté dans ASIMO (implémenté notamment dans HARK). Dans [78], Takahashi *et al.* proposent une modification dans le module de séparation de sources géométrique utilisé dans HARK et implémenté dans le robot HRP-2. Ils utilisent les fonctions de transfert de tête précalculées du robot dans le GSS pour estimer les matrices de séparation.

Conclusion

Nous avons présenté un aperçu de l'état de l'art relatif à la séparation de sources audio dans un mélange convolutif. Dans un premier temps, nous sommes intéressés aux algorithmes de séparation en les classant selon leurs critères de séparation.

Ensuite, nous nous sommes focalisés sur les algorithmes de séparation de sources qui ont été utilisés dans le cadre qui nous intéresse dans cette thèse, c'est à dire l'audition de robots. Dans les deux chapitres suivants, nous nous intéressons à deux types d'algorithmes de séparation de sources : des algorithmes basés sur l'information spatiale des sources et les algorithmes basés sur l'information structurelle des sources (la parcimonie des signaux sources dans le domaine temps-fréquence).

Deuxième partie

Séparation de sources basée sur l'information spatiale et structurelle des signaux

Chapitre 3

Formation de voies : exploitation de l'information spatiale des sources

Introduction

Dans ce chapitre nous présentons une classe de méthodes de séparation de sources basée sur l'information spatiale des sources audio : la *formation de voies* (*beamforming*). Nous nous intéressons à ce type de méthodes de séparation géométrique de sources afin de l'exploiter dans nos algorithmes de séparation comme une étape de prétraitement avant l'étape de séparation basée sur l'information structurelle des sources.

Le terme *formation de voies* a été dérivé du fait que les premiers filtres spatiaux ont été développés pour former des *lobes* permettant de recevoir un signal émis d'une certaine direction et atténuant les signaux émis des autres directions. Former des voies semble indiquer la radiation de l'énergie, cependant, la formation de voies est appliquée à la radiation ou à la réception de l'énergie. Dans nos travaux, nous considérons la formation de voies pour la réception. La formation de voies a été étudiée dans plusieurs domaines comme le radar, la sismologie et les communications. Elle peut être utilisée pour différentes applications telles que la détection de présence d'un signal, l'estimation des directions d'arrivée des sources, l'amélioration de la qualité d'un signal, l'estimation d'un signal venant d'une direction donnée en la présence de bruit et de signaux interférents... Ceci est possible en construisant le *diagramme de directivité* adéquat, celui qui donne la bonne répartition de l'énergie en réception en fonction de la fréquence et des angles d'arrivées éventuels de la source audio émettrice, et ceci par rapport à l'application souhaitée.

La formation de voies est formulée comme un filtre spatial qui opère sur les sorties

d'un réseau de capteurs dans le but de former le diagramme de directivité désiré [87]. Typiquement, un filtre de formation de voies (*beamformer*) combine linéairement les séries temporelles échantillonnées spatialement par le réseau de capteurs afin d'obtenir à la sortie des séries temporelles scalaires : c'est le *filtrage spatio-temporel*. Le filtre de formation de voies idéal doit avoir un diagramme de directivité égale à 1 pour la direction d'arrivée choisie et nul dans les autres directions. Dans la pratique, ce genre de filtre ne peut être estimé : le lobe principal est toujours de largeur non nulle et les lobes secondaires sont de hauteur non nulle.

Dans le cadre de nos travaux, nous nous sommes intéressés à la formation de voies comme méthode de séparation de sources selon leurs *directions d'arrivées* (DOA : Directions of Arrival). La direction d'arrivée d'une source est l'angle que fait la perpendiculaire à son front d'onde avec le réseau de capteurs. Les méthodes de formation de voies peuvent être classées en deux catégories, selon leur dépendance aux données reçues :

- la formation de voies indépendante des données observés, appelée aussi *formation de voies fixe* ;
- la formation de voies dépendante des donnés observés, appelée aussi *formation de voies adaptative*.

Dans la suite, nous présenterons dans un premier temps le principe de la formation de voies. Nous nous intéresserons ensuite à la formation de voies adaptative et à la formation fixe en soulignant les avantages et les inconvénients de chacune de ces deux méthodes. Ensuite, nous introduirons notre méthode de formation de voies fixe avec les fonctions de transfert de tête (HRTF).

3.1 Formation de voies : principe

Supposons que l'on dispose d'un réseau de M capteurs de géométrie quelconque. Une onde plane $s\left(t\right)$ issue d'une source audio située en champ lointain arrive sur le réseau de capteurs avec une direction d'arrivée θ. Les sorties du réseau de capteurs sont notées $\mathbf{x}\left(t\right) = \left[x_1\left(t\right), \ldots, x_M\left(t\right)\right]^T$, t étant l'indice de temps, et sont modélisées comme la convolution de la source $s(t)$ et les réponses impulsionnelles des différents chemins de propagation entre cette source et les capteurs. Ces réponses impulsionnelles sont tronquées à la longueur L et s'écrivent en fonction de la direction d'arrivée θ de la source $s(t)$ comme $\mathbf{h}\left(l,\theta\right) = \left[h_1(l,\theta), \ldots, h_M(l,\theta)\right]^T$, avec $0 \leq l \leq L-1$. Le signal à la sortie du capteur m s'écrit :

$$x_m(t) = \sum_{l=0}^{L-1} h_m(l,\theta)\, s(t-l) \tag{3.1}$$

Nous voulons estimer un filtre de formation de voies qui nous permettra de rehausser le niveau de ce signal et de rejeter les signaux interférents arrivant de directions autre que θ. Ceci est possible en construisant un lobe qui vise la direction d'arrivée désirée θ.

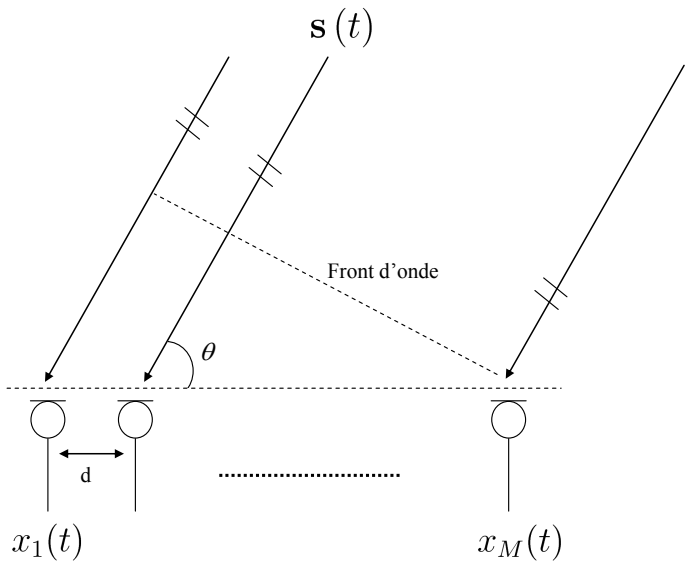


FIGURE 3.1 – Illustration d'un front d'onde plane arrivant sur un réseau de capteurs linéaire

Soit le filtre de formation de voies dans le domaine temporel $\underline{\mathbf{b}}(l,\theta) = [\underline{b}_1(l,\theta), \ldots, \underline{b}_M(l,\theta)]^T$ relatif à une direction de visée θ, avec $1 \leq l \leq L'$ et L' est la longueur de ce filtre. Les signaux à la sortie de la formation de voies sont exprimés dans le domaine temporel comme suit :

$$z(t) = \sum_{m=1}^{M} \sum_{l=0}^{L-1} \underline{b}_m(l,\theta)\, x_m(t-l) = \sum_{l=0}^{L-1} \underline{\mathbf{b}}^T(l,\theta)\, \mathbf{x}(t-l) \tag{3.2}$$

Les filtres spatiaux de formation de voies ont été développés pour les signaux à bande étroite qui peuvent être caractérisés par une fréquence unique. Pour les signaux large bande comme la parole, une transformation dans le domaine fréquentiel est nécessaire et les filtres de formation de voies à bande étroite sont développés pour

chaque fréquence indépendamment. Par conséquent, l'équation (3.2) s'écrit dans le domaine temps-fréquence :

$$z\left(f,k\right)=\mathbf{b}^{H}\left(f,\theta\right)\mathbf{X}\left(f,k\right) \tag{3.3}$$

où $\mathbf{b}^{H}\left(f,\theta\right)=\left[b_{1}\left(f,\theta\right),\ldots,b_{M}\left(f,\theta\right)\right]^{T}$ et $\mathbf{X}\left(f,k\right)\simeq\mathbf{H}\left(f\right)S\left(f,k\right)$ avec $\mathbf{X}(f,k)=\left[X_{1}(f,k),\ldots,X_{M}(f,k)\right]^{H}$ (respectivement $S(f,k)$) est la TFCT de$\{\mathbf{x}\left(t\right)\}_{1\leq t\leq T}$ (respectivement de $\{s\left(t\right)\}_{1\leq t\leq T}$) à la fréquence $f\in\left[1,\frac{N_{f}}{2}+1\right]$ et la fenêtre d'analyse $k\in[1,N_{T}]$. $\mathbf{H}(f,\theta)$ est la transformée de Fourier des filtres de mélanges $\{\mathbf{h}\left(l,\theta\right)\}_{0\leq l\leq L-1}$ de dimension $M\times1$.

Nous voulons que le diagramme de directivité, défini comme la magnitude au carré de la *réponse directionnelle* de cette formation de voies, pointe vers la direction d'arrivée θ. La réponse directionnelle de la formation de voies s'écrit en fonction du *vecteur directionnel* $\mathbf{d}\left(f,\theta\right)$ comme suit :

$$r\left(f,\theta\right)=\mathbf{b}^{H}\left(f,\theta\right)\mathbf{d}\left(f,\theta\right) \tag{3.4}$$

Le vecteur directionnel est la réponse en fréquence du réseau de capteurs, il représente les délais de phase d'une onde plane calculés aux niveaux des éléments du réseau de capteurs. Cette réponse est une fonction de l'angle de visée θ et de la configuration du réseau. Dans un champ libre, le vecteur directionnel d'un réseau de capteurs s'écrit comme suit :

$$\mathbf{d}\left(f,\theta\right)=\left[e^{-j2\pi f\mathcal{F}_{1}(\tau(\theta))},\ldots,e^{-j2\pi f\mathcal{F}_{M}(\tau(\theta))}\right] \tag{3.5}$$

où $\tau\left(\theta\right)$ est la *différence de temps d'arrivée* (TDOA : Time Difference Of Arrival) entre le premier et le deuxième capteur d'une onde plane arrivant d'une direction d'arrivée égale à θ, et $\mathcal{F}_{m}\left(\tau\left(\theta\right)\right)$ est le délai temporel relatif entre le 1^{er} et le $m^{ième}$ capteur. Pour un réseau de capteurs linéaire uniforme, $\mathcal{F}_{m}\left(\tau\left(\theta\right)\right)=\left(m-1\right)d\cos\left(\theta\right)/c$ où d est la distance entre deux capteurs et c est la vitesse du son (*cf.* figure 3.1).

L'estimation du filtre de formation de voies $\mathbf{b}\left(f,\theta\right)$ pour la fréquence f et la direction d'arrivée θ peut se faire d'une manière adaptative en fonction des données observées, ou d'une manière fixe, complètement indépendante des observations. C'est le sujet des deux prochaines sections.

3.2 Formation de voies adaptative

Dans une formation de voies statistiquement optimale, les poids du filtre sont estimés en se basant sur les statistiques des signaux reçus aux capteurs. Le but est d'optimiser la réponse directionnelle de la formation de voies de telle sorte que sa sortie contienne une contribution minimale des signaux provenant d'autres directions que la direction désirée [87]. Dans la suite, nous présentons deux exemples de ces formations de voies adaptatives calculés dans le domaine fréquentiel [15, 87] : la formation de voies MVDR ou Capon et la formation de voies à maximisation du rapport signal sur sources.

3.2.1 Capon ou MVDR

La méthode MVDR (Minimum Variance Distortionless Response) de Capon est sans doute la méthode de formation de voies adaptative la plus utilisée. Le principe de MVDR est d'estimer les coefficients du filtre $\mathbf{b}\left(f,\theta\right)$ qui minimise l'énergie du signal à la sortie $\mathbb{E}\left(z^2\left(f,k\right)\right) = \mathbf{b}^H(f,\theta)\mathbf{R}_{xx}(f,k)\mathbf{b}(f,\theta)$, avec la contrainte que le signal désiré $S(f,k)$ ne soit pas affecté. Le problème MVDR s'écrit :

$$\min_{\mathbf{b}(f,\theta)} \mathbf{b}^H(f,\theta)\mathbf{R}_{xx}(f,k)\mathbf{b}(f,\theta) \text{ tel que } \mathbf{b}^H(f,\theta)\mathbf{H}\left(f,\theta\right) = 1 \tag{3.6}$$

où $\mathbf{R}_{xx}\left(f,k\right) = \frac{1}{M}\sum_{m=1}^{M} X_m^H\left(f,k\right) X_m\left(f,k\right)$. La solution de l'équation précédente donne le filtre de formation de voies par la méthode MVDR et pour la direction de visée θ :

$$\mathbf{b}(f,\theta) = \frac{\mathbf{R}_{xx}^{-1}(f,k)\mathbf{H}\left(f,\theta\right)}{\mathbf{H}\left(f,\theta\right)^H \mathbf{R}_{xx}^{-1}(f,k)\mathbf{H}\left(f,\theta\right)} \tag{3.7}$$

3.2.2 Maximisation du rapport signal sur bruit

Comme nous le verrons dans la section suivante, la formation de voies fixe exploite la géométrie du réseau de capteurs pour optimiser son diagramme de directivité. Cependant, la capacité d'une formation de voies fixe à supprimer le bruit et les sources interférentes est limitée par plusieurs facteurs comme l'ouverture géométrique du réseau de capteurs. Une manière d'obtenir un rapport signal sur bruit (SNR : Signal to Noise Ratio) supérieur quand la géométrie du réseau de capteurs est fixée passe par l'utilisation des caractéristiques des signaux source et bruit. Dans ce cas, on cherche à estimer le filtre optimal qui maximise le SNR à la sortie de la formation de voies.

Nous considérons le même modèle que dans le paragraphe précédent, avec l'ajout d'un vecteur de bruit $\mathbf{v}(t) = [v_1(t), \ldots, v_M(t)]^T$. Le signal et le bruit sont supposés être décorrélés, la matrice d'autocorrélation du vecteur $\mathbf{X}(f,k)$ s'écrit : $\mathbf{R}_{xx}(f,k) = \mathbf{H}(f)^H \mathbf{R}_{ss}(f,k)\mathbf{H}(f,\theta) + \mathbf{R}_{vv}(f,k)$, où $\mathbf{R}_{vv}(f,k)$ est la matrice d'autocorrélation du bruit $\mathbf{V}(f,k)$ dans le domaine temps-fréquence. Le SNR de sortie s'écrit à la fréquence f comme :

$$SNR(f) = \frac{\mathbf{b}^H(f,\theta)\,\mathbf{H}^H(f,\theta)\,\mathbf{R}_{ss}(f,k)\mathbf{H}(f,\theta)\,\mathbf{b}(f,\theta)}{\mathbf{b}^H(f,\theta)\,\mathbf{R}_{vv}(f,k)\,\mathbf{b}(f,\theta)} \tag{3.8}$$

Dans un traitement par réseau de capteurs, nous espérons supprimer le plus de bruit possible. Une méthode possible pour réaliser ceci est d'estimer le filtre $\mathbf{b}(f,\theta)$ qui maximise $SNR(f)$, la solution est le vecteur propre associé à la plus grande valeur propre de $\mathbf{R}_{vv}^{-1}(f,k)\,\mathbf{R}_{ss}(f,k)$, avec $\mathbf{H}(f,\theta)$ égal au vecteur directionnel du réseau de capteurs.

3.3 Formation de voies fixe

Dans une formation de voies fixe, les poids des filtres spatio-temporels $\mathbf{b}(f,\theta)$ sont estimés de telle sorte que la réponse directionnelle $r(f,\theta) = \mathbf{b}^H(f,\theta)\,\mathbf{d}(f,\theta)$ de la formation de voies soit approximée à une réponse désirée $r_d(f,\theta)$, indépendamment des données observés à la sortie des capteurs. Ceci est similaire à la construction de filtres RIF à partir d'un gabarit.

Par analogie avec les techniques employées pour la construction des filtres RIF, les poids des filtres de formation de voies $\mathbf{b}(f,\theta)$ sont ceux qui minimisent la norme l_p pondérée de la différence entre la réponse directionnelle effective et la réponse directionnelle désirée. Nous considérons la technique la plus utilisée qui est la méthode des moindres carrés. Pour estimer les filtres de formation de voies fixe avec l'optimisation de la norme l_2, nous considérons la minimisation du carré de l'erreur entre la réponse directionnelle effective et la réponse directionnelle désirée [15] :

$$\min_{\mathbf{b}(f,\theta)} \sum_{\theta\in\Theta} \mathcal{V}(\theta)\,|r(f,\theta) - r_d(f,\theta)|^2 \tag{3.9}$$

Θ étant un ensemble d'angles couvrant l'espace où se trouve la source qui nous intéresse (exemple : $\Theta = [0,\pi]$) et $\mathcal{V}(\theta)$ est une fonction de poids positive pour accentuer ou diminuer l'importance de certains angles. En remplaçant l'expression de la réponse directionnelle (3.4) dans l'équation (3.9), l'erreur entre cette réponse

directionnelle et la réponse directionnelle désirée à la fréquence f s'écrit :

$$\epsilon^2(f) = \mathbf{b}^H(f, \theta \in \Theta)\,\mathbf{Q}(f)\,\mathbf{b}(f, \theta \in \Theta) - 2\mathbf{b}^H(f, \theta \in \Theta)\,\mathbf{p}(f) + \sum_{\theta \in \Theta} \mathcal{V}(\theta)\,|r_d(f, \theta)|^2 \tag{3.10}$$

où :

$$\mathbf{Q}(f) = \sum_{\theta \in \Theta} \mathcal{V}(\theta)\,\mathbf{d}(f, \theta)\,\mathbf{d}^H(f, \theta) \tag{3.11}$$

$$\mathbf{p}(f) = \sum_{\theta \in \Theta} \mathcal{V}(\theta)\,\mathrm{Re}\left[\mathbf{d}(f, \theta)\,r_d(f, \theta)\right] \tag{3.12}$$

Re(.) étant la partie réelle.

Faire la différentielle de ϵ^2 par rapport à $\mathbf{b}(f, \theta)$ et mettre le résultat à zéro donne l'expression du filtre de formation de voies selon la méthode des moindres carrés :

$$\mathbf{b}_{MC}(f, \theta \in \Theta) = \mathbf{Q}^{-1}(f)\,\mathbf{p}(f) \tag{3.13}$$

On remarque que la matrice $\mathbf{Q}(f)$ est fonction de $\mathcal{F}_m(\tau(\theta))$ et que le vecteur $\mathbf{p}(f)$ est fonction de $\mathcal{F}_m(\tau(\theta))$ et de $r_d(f, \theta)$, par conséquent, le filtre au sens des moindres carrés dépend de la géométrie du réseau de capteurs et de la réponse directionnelle désirée. Si nous voulons construire une formation de voies qui laisse passer le signal incident venant d'un angle entre θ_1 et θ_2, la réponse directionnelle désirée s'écrit dans ce cas :

$$r_d(f, \theta) = \begin{cases} 1 & \text{si } \theta_1 \leq \theta \leq \theta_2 \\ 0 & \text{sinon} \end{cases} \tag{3.14}$$

3.4 Les fonctions de transfert de tête (HRTF)

La localisation d'une source sonore par un humain se fait à travers deux indices binauraux [16] : la *différence de temps interaural* (ITD : Interaural Time Difference) et la *différence d'intensité interaurale* (IID : Interaural Intensity Difference). L'ITD est la différence en temps d'arrivée d'un front d'onde sonore aux oreilles droite et gauche. L'IID est la différence en l'amplitude d'un son qui atteint les oreilles droite et gauche. Il est connu que l'ITD et l'IID sont important pour la perception du son dans un plan horizontal. Cependant, si la source sonore est autorisée à varier en

élévation et distance, l'ITD et l'IID ne spécifient pas une unique position spatiale : ils sont identiques pour des sources sonores placées sur un cône nommé le cône de confusion (*cf.* figure 3.2). La localisation d'une source placée sur ce cône n'est pas possible si nous utilisons uniquement les paramètres ITD et IID.

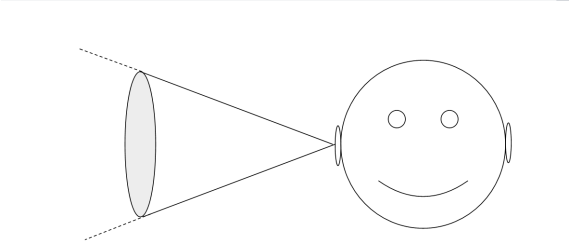


Figure 3.2 – Le cône de confusion

Cependant, le système d'audition humain est capable de définir les directions de ces sources. Ceci s'explique par le fait que pour l'audition humaine, une source sonore subit un filtrage spectral par la tête et le pavillon de l'oreille, ainsi, une fonction de transfert entre la source et chacune des deux oreilles est définie : c'est la *fonction de transfert de tête* (HRTF : Head Related Transfer Function).

Une HRTF prend en compte la différence de temps interaural, la différence d'intensité interaurale et la forme de la tête et du pavillon. La HRTF caractérise comment un son émis d'une direction spécifique et altéré par la tête et le pavillon est reçu à l'oreille. La notion de HRTF reste la même si on remplace la tête humaine par la tête d'un mannequin de vitrine (dummy) et les deux oreilles par deux microphones. Nous gardons toujours la même notion de HRTF si nous augmentons le nombre de capteurs à plus que 2. Avoir plus que deux capteurs fixés autour d'une tête permet de capturer plus précisément l'effet de cette tête sur le champ sonore environnant.

Nous étendons le concept habituel de HRTF binaurales au contexte de l'audition des robots où un humanoïde est équipé d'un réseau de microphones (plus de deux capteurs). Dans notre cas, la HRTF $h_m(f, \theta)$ à la fréquence f caractérise comment un signal émis d'une direction spécifique θ est reçu au $m^{ième}$ microphone fixé à la tête. La réponse impulsionnelle qui correspond à la représentation temporelle de la HRTF s'appelle *réponse impulsionnelle de tête* (HRIR : Head Related Impulse Response), un exemple de HRIR est montré à la figure 3.3.

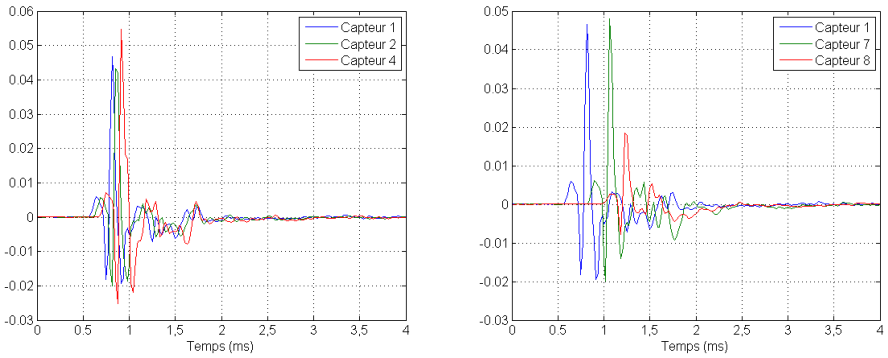


FIGURE 3.3 – Les réponses impulsionnelles de têtes (HRIR : Head Related Impulse Response) : la représentation temporelle des HRTF pour différents micros de la tête du mannequin de vitrine

3.5 Formation de voies fixe en utilisant les HRTF

3.5.1 Vers la modélisation de la variété du réseau de capteurs

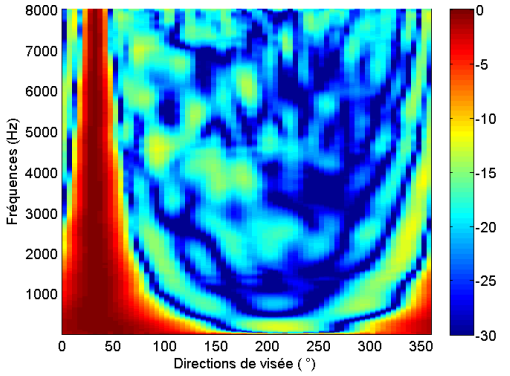
Dans le cas de l'audition des robots, la géométrie du réseau de microphones est fixée une fois pour toute. Une fois la géométrie du réseau de capteurs connue et les directions de visée choisies par une technique de localisation ou autre, la formation de voies fixe utilise ces informations spatiales pour construire le diagramme de directivité désiré. Les caractéristiques du diagramme de directivité (largeur des lobes, amplitudes des lobes secondaires et la position des zéros) sont donc fixées pour tous les scénarios [1] et calculées une seule fois, indépendamment des mélanges mesurés aux capteurs.

Afin d'estimer les filtres de formation de voies qui donneront le diagramme de directivité voulu, la technique des moindres carrés introduite dans la section précédente est utilisée. Mais pour procéder à l'estimation de ces filtres, nous avons besoin de calculer les vecteurs directionnels relatif à la géométrie du réseau de capteurs. Dans le cas de l'audition des robots, les capteurs sont souvent fixés autour de la tête du robot. Le modèle en champ libre du vecteur directionnel présenté dans l'équation (3.5) ne prend pas en compte l'influence de la tête sur le champ acoustique environnant, et dans ce cas, la variété du réseau de capteurs n'est pas modélisée (inconnue).

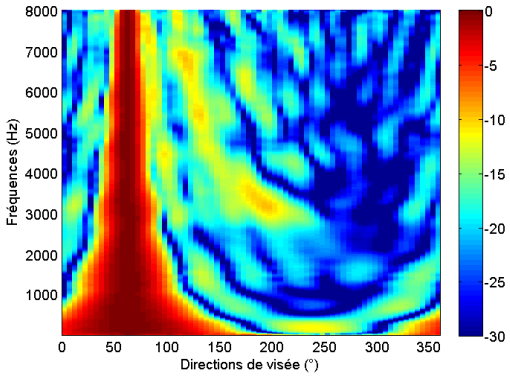
Afin de modéliser la variété du réseau de capteurs, nous proposons d'utiliser les HRTF comme vecteurs directionnels pour l'estimation des filtres de formation

1. Ici, le mot « scénario » fait référence aux différentes directions d'arrivée que l'ont peut avoir dans notre problème de séparation ainsi qu'au changement du nombre des sources.

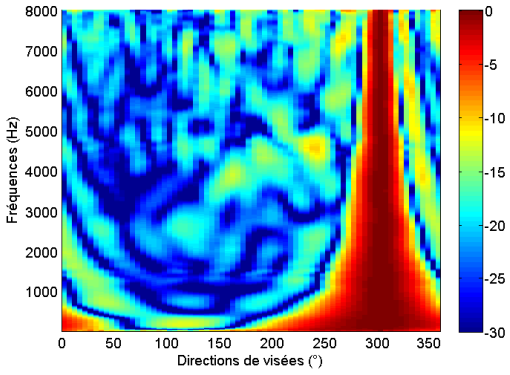
de voies et remplacer la variété inconnue du réseau de capteurs par une distribution discrète de HRTF d'un groupe de N_s directions de visée choisi *a priori* $\Theta = [\theta_1, \ldots, \theta_{N_S}]$. Les HRTF sont mesurées dans une chambre anéchoïque comme ça sera expliqué à la section 7.6.



(a) Angle de visée $\theta = 30°$



(b) Angle de visée $\theta = 60°$



(c) Angle de visée $\theta = 300°$

FIGURE 3.4 – La réponse directionnelle $r_{\text{hrtf}}(f, \theta) = \mathbf{b}^H(f, \theta)\, \mathbf{d}_{\text{hrtf}}(f, \theta)$ (en dB) relative à différents angles de visée, construite à partir des HRTF

3.5.2 Estimation des filtres de formation de voies par les HRTF

Nous proposons d'utiliser les HRTF comme vecteurs directionnels $\{\mathbf{d}(f,\theta)\}_{\theta\in\Theta}$, où $\Theta = \{\theta_1, \ldots, \theta_{N_S}\}$ est un groupe de N_S directions de visée choisi *a priori*. Soit $h_m(f,\theta)$ une fonction de transfert de tête d'un point d'émission localisé à θ jusqu'au $m^{ième}$ capteur, à la fréquence f. Le vecteur directionnel est donc :

$$\mathbf{d}_{\text{hrtf}}(f,\theta) = [h_1(f,\theta), \ldots, h_M(f,\theta)]^T \tag{3.15}$$

Nous voulons une formation de voies qui extrait le signal émis de la direction θ_i. Dans ce cas, la réponse directionnelle désirée s'écrit :

$$r_d(f,\theta) = \begin{cases} 1 & \text{si } \theta = \theta_i \\ 0 & \text{sinon} \end{cases} \tag{3.16}$$

Étant données l'équation (3.15) et la réponse directionnelle (3.16), le filtre de formation de voies selon la méthode des moindres carrés (*cf.* équation (3.13)) et pour une direction de visée θ_i s'écrit :

$$\mathbf{b}(f,\theta_i) = \mathbf{R}_{\mathbf{dd}}^{-1}(f)\,\mathbf{d}_{\text{hrtf}}(f,\theta_i) \tag{3.17}$$

où $\mathbf{R}_{\mathbf{dd}}(f) = \frac{1}{N_S}\sum_{\theta\in\Theta}\mathbf{d}_{\text{hrtf}}(f,\theta)\,\mathbf{d}_{\text{hrtf}}^H(f,\theta)$. Nous considérons la version normalisée de ce filtre :

$$\mathbf{b}(f,\theta_i) = \frac{\mathbf{R}_{\mathbf{dd}}^{-1}(f)\,\mathbf{d}_{\text{hrtf}}(f,\theta_i)}{\mathbf{d}_{\text{hrtf}}^H(f,\theta_i)\,\mathbf{R}_{\mathbf{dd}}^{-1}(f)\,\mathbf{d}_{\text{hrtf}}(f,\theta_i)} \tag{3.18}$$

Si nous voulons choisir un sous-enemble de K directions de visée $\theta_1, \ldots, \theta_K$ à partir des N_S directions de visée pour lequelles nous avons calculé les filtres de formation de voies, la matrice de formation de voies $\mathbf{B}(f)$ s'écrit comme suit :

$$\mathbf{B}(f) = [\mathbf{b}(f,\theta_1), \ldots, \mathbf{b}(f,\theta_K)]^T \tag{3.19}$$

La figure 3.4 montre des exemples de réponses directionnelles construites à partir des HRTF.

Conclusion

Nous avons présenté dans ce chapitre une classe de méthodes de séparation de sources : la formation de voies. Après une introduction au principe de la formation de voies, nous avons consacré la deuxième section de ce chapitre à la formation de voies adaptative qui utilise les statistiques des données aux capteurs et nous nous sommes intéressés dans la troisième section à la formation de voies fixe, indépendante des données reçues. C'est ce dernier type de formation de voies qui nous intéresse dans nos travaux. Nous avons étendu le principe de formation de voies fixe à l'audition des robots en tenant compte de l'effet de la tête de celui-ci sur le champ sonore proche et ceci en modélisant la variété du réseau de capteurs par les fonctions de transfert de tête (HRTF). Cette nouvelle méthode de séparation géométrique basée sur les HRTF constitue un pilier de nos travaux sur la séparation de sources et sera utilisée comme un prétraitement à l'étape de séparation de sources dans nos algorithmes de séparation à deux étapes.

Chapitre 4

Séparation aveugle de sources audio basée sur l'information structurelle des sources

Introduction

Par opposition à la séparation de sources par formation de voies que nous avons vue dans le chapitre précédent et qui se base essentiellement sur l'information spatiale des sources, nous présentons dans ce chapitre des algorithmes de séparation basés sur l'information *structurelle* des sources audio :

1. l'indépendance des sources en utilisant l'analyse en composantes indépendantes (ACI) de Comon [23] ;
2. la parcimonie des sources en utilisant la minimisation de la norme l_p, avec avec $0 < p \leq 1$.

Nous nous sommes intéressés en particulier à la séparation des mélanges convolutifs basée sur la parcimonie des signaux sources dans le domaine temps-fréquence : les sources séparées sont les signaux les plus parcimonieux que l'on puisse obtenir à partir des mélanges reçus. Le but est donc de trouver un critère de séparation qui maximise la parcimonie des sources estimées.

Dans un premier temps, nous utilisons une mesure de parcimonie assez connue grâce à sa propriété de convexité : la norme l_1. Ensuite, nous avons développé un algorithme de séparation basé sur la pseudo-norme l_p, avec $0 < p < 1$: plus le paramètre p est proche de 0, plus la contrainte de parcimonie est dure, la norme l_1 ayant la contrainte la plus souple. Nous nous sommes penchés sur le problème de

minimisation de la pseudo-norme l_p, avec $0 < p < 1$ en développant un critère de séparation paramétré qui permet de rendre la contrainte de parcimonie de plus en plus rigide au fur et à mesure que l'algorithme avance dans ses itérations.

Nous commençons ce chapitre par l'introduction de la méthode d'optimisation du gradient naturel d'Amari [13] utilisée pour minimiser ces critères. Ensuite, nous présentons l'algorithme d'analyse en composantes indépendantes utilisé et l'algorithme basé sur la minimisation de la norme l_1, suivis de l'algorithme de la pseudo-norme l_p paramétrée.

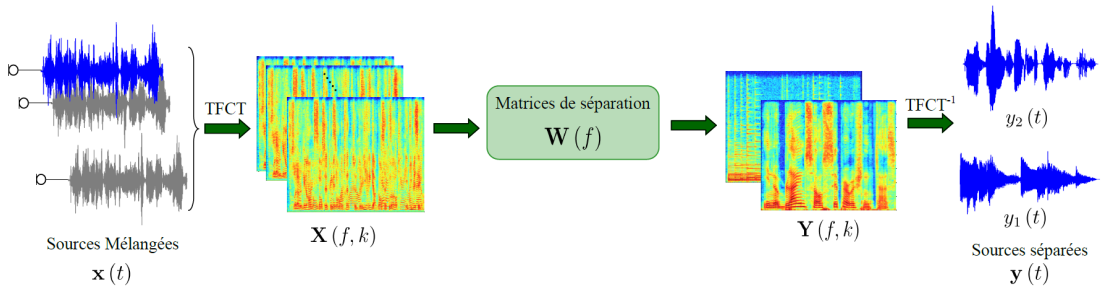


Figure 4.1 – Principe de la séparation de sources

4.1 L'algorithme d'optimisation du gradient naturel

Le gradient naturel est une méthode d'optimisation proposée par Amari et *al.* en 1996 [13]. C'est une méthode de gradient dont la direction de recherche standard du gradient est modifiée selon la structure Riemannienne locale de l'espace de paramètres. Ceci garantit l'invariance de la direction de recherche du gradient naturel par rapport aux relations statistiques entre les paramètres du modèle et conduit à une performance d'apprentissage statistiquement efficace [14].

Supposons que nous voulons mettre à jour une matrice de séparation $\mathbf{W}$ par rapport à une fonction de coût $\psi(\mathbf{W})$ (*cf.* figure 4.1). La mise à jour de cette matrice selon une méthode de descente de gradient est :

$$\mathbf{W}_{n+1} = \mathbf{W}_n - \mu \nabla \psi(\mathbf{W}_n) \tag{4.1}$$

$\nabla \psi(\mathbf{W})$ est le gradient de la fonction $\psi(\mathbf{W})$, n est un indice d'itération si nous sommes dans un cas d'optimisation itératif ou un indice de temps si nous sommes dans un cas d'optimisation adaptatif et μ est le pas de mise à jour. D'après [14], le

gradient naturel de la fonction de coût $\psi(\mathbf{W})$ est donné par :

$$\tilde{\nabla}\psi(\mathbf{W}) = \nabla\psi(\mathbf{W})\,\mathbf{W}^H\mathbf{W} \tag{4.2}$$

La mise à jour de la matrice $\mathbf{W}$ en utilisant la méthode d'optimisation du gradient naturel est donc :

$$\mathbf{W}_{n+1} = \mathbf{W}_n - \mu\nabla\psi(\mathbf{W}_n)\,\mathbf{W}_n^H\mathbf{W}_n \tag{4.3}$$

Initialisation

Pour initialiser la matrice de séparation $\mathbf{W}_0$, nous utilisons un processus de blanchiment. Le blanchiment est un prétraitement important dans la séparation de sources sur-déterminé, il permet de focaliser l'énergie du signal reçu dans l'espace signal utile. On considère $\mathbf{D}$ et $\mathbf{E}$ respectivement la matrice diagonale et la matrice unitaire de la décomposition en valeurs singulières de la matrice d'autocorrélation des données reçues $\mathbf{X}$. Soient $\mathbf{D}_M$ la matrice contenant les M premières lignes et les M première colonnes de la matrice $\mathbf{D}$ et $\mathbf{E}_{:M}$ la matrice contenant les M premières colonnes de la matrice $\mathbf{E}$. La matrice de séparation est donc initialisée comme suit :

$$\mathbf{W}_0 = \sqrt{\mathbf{D}_M^{-1}}\mathbf{E}_{:M}^H$$

Si nous nous trouvons dans un cas de séparation déterminée, l'initialisation de la matrice de séparation se fait avec la matrice identité :

$$\mathbf{W}_0 = \mathbf{I}$$

4.2 Analyse en composantes indépendantes

Si nous supposons que les sources sont temporellement indépendantes et identiquement distribuées (iid) et non gaussiennes, nous faisons référence aux méthodes d'analyse en composantes indépendantes (ICA : Independent Component Analysis). L'hypothèse fondamentale est que les sources inconnues sont statistiquement indépendantes [22, 23]. Si les sources admettent une densité de probabilité, cette hypothèse se traduit par la fait que la densité de probabilité jointe des sources $p_s(s_1(t), s_2(t), \ldots, s_N(t))$ peut être factorisée en le produit des densités de probabilité marginales $\{p_i(s_i(t))\}_{1\leq i\leq N}$:

$$p_{\mathbf{s}}\left(s_1(t), s_2(t), \ldots, s_N\left(t\right)\right) = p_1\left(s_1(t)\right) p_2\left(s_2(t)\right) \ldots p_N\left(s_N\left(t\right)\right) \tag{4.4}$$

Pour un mélange linéaire instantané, le but d'ACI est d'estimer une matrice de séparation $\mathbf{W}$ qui conduit aux sources estimées $\mathbf{y}\left(t\right) = \mathbf{W}\mathbf{x}\left(t\right)$ qui sont statistiquement indépendantes. Ce critère d'indépendance n'est pas pratique parce que non seulement il demande l'égalité de deux fonctions à variables multiples mais aussi il demande leur parfaite connaissance. Par conséquent, d'autres mesures d'indépendances peuvent être utilisées et conduisent à des critères de séparation plus réalisables. La minimisation de l'information mutuelle des sources à estimer est l'un de ces critères. Ce critère est populaire dans la séparation de sources avec analyse en composantes indépendantes pour plusieurs raisons. Premièrement, il est invariant par rapport à des transformations inversibles, en particulier il est invariable par rapport à une transformation d'échelle ce qui évite l'étape de blanchiment nécessaire dans d'autres méthodes de séparation. Deuxièmement, c'est un critère d'indépendance général et complet : il est non négatif et s'annule si et seulement si l'indépendance existe. Finalement, l'information mutuelle peut être interprétée en termes d'entropie et de divergence de Kullback-Leibler et est étroitement liée à la log vraisemblance.

L'*information mutuelle* entre les densités de probabilité jointe et marginale des sources à estimer est définie comme la divergence de Kullback-Leibler entre les densités $\prod_{i=1}^{N} p_i\left(y_i\left(t\right)\right)$ et $p_{\mathbf{y}}\left(y_1(t), y_2(t), \ldots y_N\left(t\right)\right)$:

$$I\left\{y_1(t), y_2(t), \ldots y_N\left(t\right)\right\} = -\mathbb{E}\log\frac{p_1\left(y_1(t)\right) p_2\left(y_2(t)\right) \ldots p_N\left(y_N\left(t\right)\right)}{p_{\mathbf{y}}\left(y_1(t), y_2(t), \ldots y_N\left(t\right)\right)} \tag{4.5}$$

Elle peut être écrite en fonction de l'entropie jointe $H\left\{y_1\left(t\right), \ldots, y_N\left(t\right)\right\} = -\mathbb{E}\log p_{\mathbf{y}}\left(y_1(t), y_2(t), \ldots y_N\left(t\right)\right)$ et des entropies marginales $H\left\{y_i\left(t\right)\right\} = -\mathbb{E}\log p_i\left(y_i(t)\right)$ pour $1 \leq i \leq N$:

$$I\left\{y_1(t), y_2(t), \ldots y_N\left(t\right)\right\} = \sum_{i=1}^{N} H\left\{y_i\left(t\right)\right\} - H\left\{y_1\left(t\right), \ldots, y_N\left(t\right)\right\} \tag{4.6}$$

L'entropie possède la propriété intéressante d'invariance par rapport à une transformation inversible.

Lemme Soient $\mathbf{x}$ un vecteur aléatoire et $\mathbf{y} = g(\mathbf{x})$ où g est une transformation différentiable et inversible avec un Jacobien (matrice des dérivées partielles) g'. Alors :

$$H\{\mathbf{y}\} = H\{\mathbf{x}\} + \mathbb{E}\log|\det g'(\mathbf{x})|$$

Pour un cas de séparation de sources instantané où $\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t)$, le critère de séparation peut s'écrire en fonction de la fonction de la matrice de séparation $\mathbf{W}$:

$$\psi(\mathbf{W}) = \sum_{i=1}^{N} H(y_i(t)) - \log|\det \mathbf{W}| \tag{4.7}$$

Ce critère n'est que théorique puisqu'il implique les entropies inconnues $H(y_i(t))$. Dans la pratique, ces entropies doivent être remplacées par leurs estimées. Ceci revient à remplacer les densités de probabilité des sources par leur estimées. Un estimateur populaire et bien connu est celui par noyau.

Le calcul de la mise à jour de la matrice de séparation $\mathbf{W}$ pour un cas de séparation de mélanges instantanés selon le critère de la maximisation de l'entropie est [13] :

$$\mathbf{W}_{n+1} = \mathbf{W}_n + \mu(\mathbf{I} - \mathbf{G}_n)\mathbf{W}_n \tag{4.8}$$

où $\mathbf{G}_n = \frac{1}{T}\sum_{t=1}^{T}\mathbf{f}(\mathbf{y}_n(t))\mathbf{y}_n^H(t)$ et $\mathbf{y}_n(t) = \mathbf{W}_n\mathbf{x}(t)$ est l'estimation des signaux sources au temps t et à l'itération n et $\mathbf{f}(\mathbf{y})$ est une fonction non linéaire.

Dans le domaine fréquentiel, le mélange convolutif est approximé à un mélange instantané. L'équation de mise à jour devient :

$$\mathbf{W}_{n+1}(f) = (1+\mu)\mathbf{W}_n(f) - \mu\mathbf{G}_n(f)\mathbf{W}_n(f) \tag{4.9}$$

où $\mathbf{G}_n(f) = \frac{1}{N_T}\sum_{k=1}^{N_T}\mathbf{f}(\mathbf{Y}_n(f,k))\mathbf{Y}_n^H(f,k)$, $\mathbf{Y}_n(f,k) = \mathbf{W}_n(f)\mathbf{X}(f,k)$ et μ est le pas de mise à jour.

La convergence du gradient naturel est conditionnée par l'initialisation de la matrice de séparation $\mathbf{W}_0(f)$ et par le choix du pas de mise à jour μ et il est un peu difficile de choisir les bons paramètres qui permettront une convergence rapide sans risquer la divergence. Douglas et Gupta [29] proposent d'imposer une contrainte d'échelle à la matrice de séparation $\mathbf{W}(f)$ afin de maintenir une amplitude constante du gradient au cours des itérations de l'algorithme. Ils affirment qu'avec cette contrainte d'échelle et un pas de mise à jour fixe μ, l'algorithme a une convergence rapide et une bonne performance indépendamment de l'amplitude de

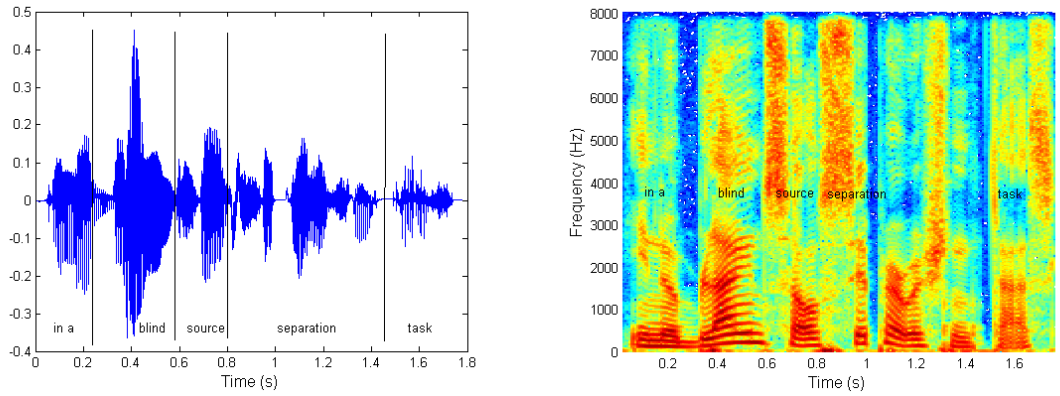
$\mathbf{X}(f,:)$ et $\mathbf{W}_0(f)$. L'équation de mise à jour devient :

$$\mathbf{W}_{n+1}(f) = (1+\mu)\, c_n \mathbf{W}_n(f) - \mu c_n^2 \mathbf{G}_n(f)\, \mathbf{W}_n(f) \tag{4.10}$$

où $c_n = \frac{1}{d_n}$ et $d_n = \frac{1}{N}\sum_{i,j=1}^{N} |G_n^{ij}(f)|$ et $G_n^{ij} = [\mathbf{G}_n(f)]_{ij}$ le $ij^{ième}$ coefficient de la matrice $\mathbf{G}_n(f)$.

4.3 Minimisation de la norme l_1

Les signaux de parole sont connus pour être parcimonieux dans le domaine temps-fréquence : le nombre de points temps-fréquence où le signal de parole est actif (c'est à dire d'énergie non négligeable) est petit par rapport au nombre total de points temps-fréquence (*cf.* figure 4.2).



(a) (b)

Figure 4.2 – Parcimonie du signal de parole dans le domaine temps fréquence (b) comparé au domaine temporel (a)

Nous considérons un critère de séparation basé sur la parcimonie des signaux dans le domaine temps-fréquence : pour chaque fréquence f, nous cherchons une matrice de séparation $\mathbf{W}(f)$ qui nous conduira aux sources estimées les plus parcimonieuses $\mathbf{Y}(f,:) = [\mathbf{Y}(f,1), \ldots, \mathbf{Y}(f,N_T)]$. De la même façon, nous définissions la matrice de mélange à chaque fréquence $\mathbf{X}(f,:) = [\mathbf{X}(f,1), \ldots, \mathbf{X}(f,N_T)]$.

Pour mesurer la parcimonie d'un signal, la norme l_1 est la mesure de parcimonie la plus utilisée grâce à sa propriété de convexité [40]. Plus un signal est parcimonieux, plus sa norme l_1 est petite. Nous proposons d'utiliser la norme l_1 comme mesure de parcimonie des signaux estimés $\mathbf{Y}(f,:)$ dans la fonction de coût suivante :

$$\psi\left(\mathbf{W}\left(f\right)\right)=\sum_{i=1}^{N}\sum_{k=1}^{N_T}\left|Y_i\left(f,k\right)\right| \tag{4.11}$$

Le but est d'estimer $\mathbf{W}(f)$ qui minimise la fonction de coût $\psi\left(\mathbf{W}\left(f\right)\right)$ et donc minimise la parcimonie des signaux sources estimés $\mathbf{Y}\left(f,:\right)$:

$$\min_{\mathbf{W}}\psi\left(\mathbf{W}\left(f\right)\right)\text{ telle que }\left\|\mathbf{W}(f)\right\|=1 \tag{4.12}$$

où $\|.\|$ est une norme matricielle. Nous minimisons la fonction de coût $\psi\left(\mathbf{W}\left(f\right)\right)$ en utilisant la méthode d'optimisation du gradient naturel présenté dans la section 4.1 de ce chapitre, la fonction de mise à jour de la matrice de séparation $\mathbf{W}(f)$ s'écrit donc :

$$\mathbf{W}_{t+1}\left(f\right)=\mathbf{W}_t\left(f\right)-\mu\nabla\psi\left(\mathbf{W}_t\left(f\right)\right)\mathbf{W}_t^H\left(f\right)\mathbf{W}_t\left(f\right) \tag{4.13}$$

Calculons l'expression de l'équation de mise à jour. La différentielle de $\psi\left(\mathbf{W}\left(f\right)\right)$ est :

$$\mathrm{d}\psi\left(\mathbf{W}\left(f\right)\right)=\mathbf{f}\left(\mathbf{Y}\left(f,:\right)\right)\mathrm{d}\mathbf{Y}^T\left(f,:\right) \tag{4.14}$$

où $\mathbf{f}\left(\mathbf{Y}\left(f,:\right)\right)=\text{signe}\left(\mathbf{Y}\left(f,:\right)\right)=\left[\frac{Y(f,k)}{|Y(f,k)|}\right]_{1\leq k\leq N_T}$ est une matrice de même dimension que $\mathbf{Y}(f,:)$. Par conséquent, le gradient de $\psi\left(\mathbf{W}\right)$ s'exprime comme suit :

$$\nabla\psi\left(\mathbf{W}\left(f\right)\right)=\mathbf{f}\left(\mathbf{Y}\left(f,:\right)\right)\mathbf{X}^T\left(f,:\right) \tag{4.15}$$

ce qui implique l'expression du gradient naturel de $\psi\left(\mathbf{W}_t(f)\right)$:

$$\begin{aligned}\tilde{\nabla}\psi\left(\mathbf{W}_t\left(f\right)\right)&=\nabla\psi\left(\mathbf{W}_t\left(f\right)\right)\mathbf{W}_t^H\left(f\right)\mathbf{W}_t\left(f\right)\\&=\mathbf{f}\left(\mathbf{Y}_t\left(f,:\right)\right)\mathbf{Y}_t^H\left(f,:\right)\mathbf{W}_t\left(f\right)\end{aligned} \tag{4.16}$$

Si on pose $\mathbf{G}_t\left(f\right)=\mathbf{f}\left(\mathbf{Y}_t\left(f,:\right)\right)\mathbf{Y}_t^T\left(f,:\right)$, la mise à jour de $\mathbf{W}_t\left(f\right)$ à la fréquence f s'écrit :

$$\mathbf{W}_{t+1}\left(f\right)=\mathbf{W}_t\left(f\right)-\mu\mathbf{G}_t\left(f\right)\mathbf{W}_t\left(f\right) \tag{4.17}$$

Nous utilisons la mise à jour avec la contrainte d'échelle présentée dans la section précédente [29] :

$$\mathbf{W}_{t+1}\left(f\right)=c_t\left(f\right)\mathbf{W}_t\left(f\right)-\mu c_t^2\left(f\right)\mathbf{G}_t\left(f\right)\mathbf{W}_t\left(f\right) \tag{4.18}$$

où $c_t(f) = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|g_t^{ij}(f)\right|}$ et $g_t^{ij}(f) = [\mathbf{G}_t(f)]_{ij}$. L'algorithme 4.1 résume les étapes de cette méthode de séparation par minimisation de la norme l_1.

Algorithme 4.1 Algorithme de la minimisation de la norme l_1

1. *Entrée* : les sorties du réseau de capteurs $\mathbf{x} = [\mathbf{x}(t_1), \ldots, \mathbf{x}(t_T)]$, le nombre de sources N et le pas d'optimisation μ
2. $\{\mathbf{X}(f,k)\}_{\substack{1 \leq f \leq N_f \\ 1 \leq k \leq N_T}} = \mathrm{TFCT}(\mathbf{x})$
3. pour chaque fréquence f,
   (a) initialiser la matrice de séparation $\mathbf{W}_0(f)$ par un processus de blanchiment
   (b) $\mathbf{Y}_0(f,:) = \mathbf{W}_0(f)\,\mathbf{X}(f,:)$
   (c) pour chaque itération t,
      i. $\mathbf{f}(\mathbf{Y}_t(f,:)) = \mathrm{sign}(\mathbf{Y}_t(f,:))$
      ii. $\mathbf{G}_t(f) = \mathbf{f}(\mathbf{Y}_t(f,:))\,\mathbf{Y}_t^T(f,:)$
      iii. $c_t(f) = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|g_t^{ij}(f)\right|}$
      iv. $\mathbf{W}_{t+1}(f) = c_t(f)\,\mathbf{W}_t(f) - \mu c_t^2(f)\,\mathbf{G}_t(f)\,\mathbf{W}_t(f)$
      v. $\mathbf{Y}_{t+1}(f,:) = \mathbf{W}_{t+1}(f)\,\mathbf{X}(f,:)$
4. Résolution du problème de permutation
5. *Sorties* : Les sources estimées $\mathbf{y} = \mathrm{TFCTI}\left(\{\mathbf{Y}(f,k)\}_{\substack{1 \leq f \leq N_f \\ 1 \leq k \leq N_T}}\right)$

4.4 Minimisation de la pseudo-norme l_p paramétrée

4.4.1 Principe

Nous avons présenté dans la section précédente un critère de séparation basé sur la minimisation de la norme l_1 des sources à estimer. D'une manière plus générale, la pseudo-norme l_p, $0 < p \leq 1$ est une mesure de parcimonie dont la *dureté* dépend du paramètre p. Soit le signal $\mathbf{x} = [x_1, \ldots, x_N]$, sa pseudo-norme l_p est définie par $\|\mathbf{x}\|_p = |\sum_k |x_k|^p|^{\frac{1}{p}}$. La mesure de parcimonie de $\mathbf{x}$ avec la pseudo-norme l_p est d'autant plus dure que p est proche de 0 : l'exemple le plus extrême étant la pseudo-norme l_0 avec $\|\mathbf{x}\|_0$ le nombre d'échantillons non nul de $\mathbf{x}$. La fonction de coût devient en utilisant la pseudo-norme l_p avec $0 < p \leq 1$:

$$\psi\left(\mathbf{W}(f)\right) = \left| \sum_{i=1}^{N} \sum_{k=1}^{N_T} \left|Y_i\left(f,k\right)\right|^p \right|^{\frac{1}{p}} \tag{4.19}$$

Nous n'avons aucune connaissance *a priori* sur le choix du paramètre p pour une tâche de séparation de sources, à part que le problème devient non convexe (concave) pour un $0 < p < 1$. Comme nous voulons atteindre l'état le plus parcimonieux possible des sources estimées dans le domaine temps-fréquence, nous proposons de *durcir* la contrainte de parcimonie au fur et à mesure que nous avançons dans les itérations de l'algorithme d'optimisation. L'idée est de faire décroître le paramètre p de la contrainte la moins dure, $p = 1$, à la contrainte la plus dure $p \simeq 0$. Cependant, changer la pseudo-norme l_p au cours des itérations de l'algorithme peut conduire à une divergence, nous proposons donc de décroître le paramètre p selon une fonction sigmoïde avec un pas d'avancement assez petit, de telle sorte que la convergence de l'algorithme ne soit pas perturbée. (*cf.* figure 4.3). La fonction de coût devient donc [5] :

$$\hat{\psi}\left(\mathbf{W}(f)\right) = \left| \sum_{i=1}^{N} \sum_{k=1}^{N_T} \left|Y_i\left(k,f\right)\right|^{p(t)} \right|^{\frac{1}{p(t)}} \tag{4.20}$$

Le paramètre $l_{p(t)}$ devient donc dépendant de l'itération t, le paramètre $p\left(t\right)$ peut par exemple s'écrire selon cette fonction sigmoïde :

$$p(t) = \frac{1}{1 - exp(-L + \frac{(t-1)2L}{K_0})} \tag{4.21}$$

où L est le rang de calcul de la sigmoïde (sa largeur) et K_0 est le nombre d'itérations de l'algorithme.

4.4.2 Algorithme proposé

Notre critère de séparation consiste à minimiser la pseudo-norme l_p des sources estimées $\mathbf{Y}(f,:)$ par rapport à la matrice de séparation $\mathbf{W}(f)$ en minimisant la fonction de coût (4.20) $\hat{\psi}\left(\mathbf{W}(f)\right)$:

$$\min_{\mathbf{W}(f)} \hat{\psi}\left(\mathbf{W}(f)\right) \text{ telle que } \|\mathbf{W}(f)\| = 1 \tag{4.22}$$

où $\|.\|$ est une norme matricielle. Afin de résoudre l'équation (4.22), nous utilisons la méthode d'optimisation du gradient naturel avec normalisation des coefficients. La mise à jour de la matrice de séparation $\mathbf{W}(f)$ est donnée par :

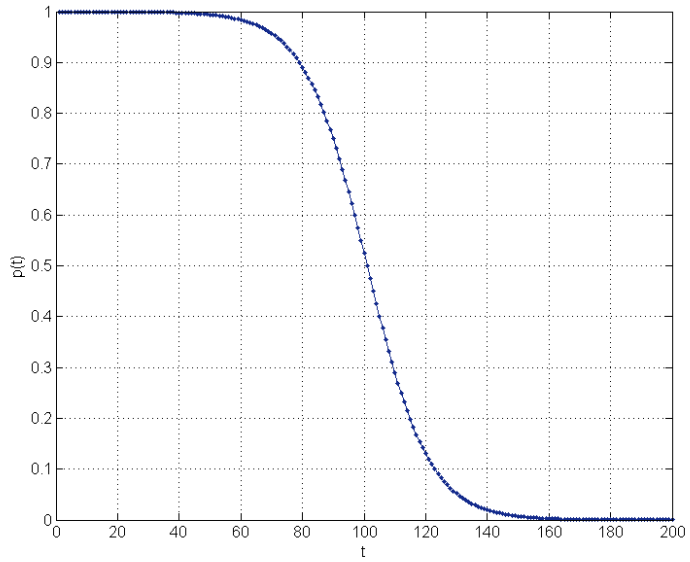


Figure 4.3 – Le paramètre p comme une fonction logistique, $p = p(t) = \frac{1}{1-exp(-L+\frac{(t-1)2L}{K_0})}$, L est le rang de calcul de la sigmoïde et $K_0 = 200$ est le nombre d'itération de l'algorithme

$$\mathbf{W}_{t+1}(f) = \mathbf{W}_t(f) - \mu\tilde{\nabla}\hat{\psi}\left(\mathbf{W}_t(f)\right) \tag{4.23}$$

où

$$\tilde{\nabla}\hat{\psi}\left(\mathbf{W}(f)\right) = \nabla\hat{\psi}\left(\mathbf{W}(f)\right)\mathbf{W}^H(f)\mathbf{W}(f) \tag{4.24}$$

est le gradient naturel de la fonction $\hat{\psi}\left(\mathbf{W}(f)\right)$ et t est un indice d'itération si nous sommes dans un cas itératif ou de temps si nous sommes dans un cas adaptatif. D'après (4.23) et (4.24), la mise à jour de $\mathbf{W}(f)$ est donnée par :

$$\begin{cases} \tilde{\mathbf{W}}_{t+1}(f) = \mathbf{W}_t(f) - \mu\nabla\hat{\psi}\left(\mathbf{W}_t(f)\right)\mathbf{W}_t^H(f)\mathbf{W}_t(f) \\ \mathbf{W}_{t+1}(f) = \frac{\tilde{\mathbf{W}}_{t+1}(f)}{\left\|\tilde{\mathbf{W}}_{t+1}(f)\right\|} \end{cases} \tag{4.25}$$

La différentielle de $\hat{\psi}\left(\mathbf{W}(f)\right)$ s'écrit :

$$\mathrm{d}\hat{\psi}\left(\mathbf{W}(f)\right) = \mathbf{f}_t\left(\mathbf{Y}(f,:)\right)\mathrm{d}\mathbf{Y}^H(f,:) \tag{4.26}$$

où $\mathbf{f}_t\left(\mathbf{Y}(f,:)\right) = p(t)\left|\mathbf{Y}(f,:)\right|^{p(t)-1} \circ \mathrm{signe}\left(\mathbf{Y}(f,:)\right)$ est une matrice de même dimension que $\mathbf{Y}(f,:)$ dans laquelle chaque $(i,j)^{ème}$ entrée est $p(t)\left|Y_i\left(f,j\right)\right|^{p(t)-1}\mathrm{signe}\left(Y_i\left(f,j\right)\right)$,

signe $(Y_i(f,j)) = \frac{Y_i(f,j)}{|Y_i(f,j)|}$ et le symbole $\circ$ fait référence au produit de Hadamard (produit composante par composante). Le gradient de $\hat{\psi}(\mathbf{W}(f))$ s'écrit comme :

$$\nabla\hat{\psi}(\mathbf{W}(f)) = \mathbf{f}_t(\mathbf{Y}(f,:))\,\mathbf{X}(f,:)^T \tag{4.27}$$

D'après (4.24) et (4.27), le gradient naturel de $\hat{\psi}(\mathbf{W}_t)$ est :

$$\tilde{\nabla}\hat{\psi}(\mathbf{W}_t(f)) = \mathbf{f}_t(\mathbf{Y}_t(f,:))\,\mathbf{Y}_t^T(f,:)\mathbf{W}_t(f) \tag{4.28}$$

La mise à jour de $\mathbf{W}(f)$ pour une fréquence f est donc :

$$\mathbf{W}_{t+1}(f) = \mathbf{W}_t(f) - \mu\mathbf{G}_t(f)\mathbf{W}_t(f) \tag{4.29}$$

où $\mathbf{G}_t(f) = \mathbf{f}_t(\mathbf{Y}_t(f,:))\,\mathbf{Y}_t^T(f,:)$. En appliquant la contrainte d'échelle présentée dans la section précédente (équation 4.18) [29], l'équation de mise à jour devient :

$$\mathbf{W}_{t+1}(f) = c_t(f)\,\mathbf{W}_t(f) - \mu c_t^2(f)\,\mathbf{G}_t(f)\,\mathbf{W}_t(f) \tag{4.30}$$

où $c_t(f) = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|g_t^{ij}(f)\right|}$ et $g_t^{ij}(f) = [\mathbf{G}_t(f)]_{ij}$.

L'algorithme proposé est résumé dans le tableau 4.2.

Algorithme 4.2 Algorithme de la minimisation de la pseudo-norme l_p paramétrée

1. *Entrée* : les sorties du réseau de capteurs $\mathbf{x} = [\mathbf{x}(t_1), \ldots, \mathbf{x}(t_T)]$, le nombre de sources N et le pas d'optimisation μ
2. $\{\mathbf{X}(f,k)\}_{\substack{1 \leq f \leq N_f \\ 1 \leq k \leq N_T}} = \mathrm{TFCT}(\mathbf{x})$
3. pour chaque fréquence f,
   (a) initialiser la matrice de séparation $\mathbf{W}_0(f)$ par un processus de blanchiment
   (b) $\mathbf{Y}_0(f,:) = \mathbf{W}_0(f)\,\mathbf{X}(f,:)$
   (c) pour chaque itération t,
      i. $\mathbf{f}(\mathbf{Y}_t(f,:)) = p(t)\,|\mathbf{Y}_t(f,:)|^{p(t)-1} \circ \mathrm{sign}(\mathbf{Y}_t(f,:))$
      ii. $\mathbf{G}_t(f) = \mathbf{f}(\mathbf{Y}_t(f,:))\,\mathbf{Y}_t^T(f,:)$
      iii. $c_t(f) = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|g_t^{ij}(f)\right|}$
      iv. $\mathbf{W}_{t+1}(f) = c_t(f)\,\mathbf{W}_t(f) - \mu c_t^2(f)\,\mathbf{G}_t(f)\,\mathbf{W}_t(f)$
      v. $\mathbf{Y}_{t+1}(f,:) = \mathbf{W}_{t+1}(f)\,\mathbf{X}(f,:)$
4. Résolution du problème de permutation
5. *Sorties* : Les sources estimées $\mathbf{y} = \mathrm{TFCTI}\left(\{\mathbf{Y}(f,k)\}_{\substack{1 \leq f \leq N_f \\ 1 \leq k \leq N_T}}\right)$

Conclusion

Dans ce chapitre, nous venons de présenter les algorithmes de séparation basés sur l'information structurelle des sources. Nous nous sommes en particulier focalisés sur les critères basés sur la parcimonie des sources que nous avons développés : la minimisation de la norme l_1 et la minimisation de la pseudo-norme l_p paramétrée selon la méthode d'optimisation du gradient naturel. Dans la suite nous présentons un algorithme de séparation de sources où l'information spatiale des sources est considérée dans une phase de prétraitement à l'algorithme de séparation basé sur l'information structurelle des sources audio.

Troisième partie

Séparation de sources à deux étapes : combinaison de formation de voies et d'algorithme de séparation de sources

Chapitre 5

Séparation de sources audio avec un prétraitement de formation de voies

Introduction

Un des principaux challenges de la séparation aveugle de sources audio est celui d'obtenir de bonnes performances de séparation dans des environnements réels réverbérants. Un prétraitement avec une formation de voies peut être une solution pour améliorer les performances des algorithmes de séparation de sources dans un milieu réverbérant. Nous rappelons que la formation de voies consiste à estimer des filtres spatiaux qui agissent sur les sorties d'un réseau de capteurs dans le but de former des lobes selon un diagramme de directivité désiré (*cf.* chapitre 3). Nous rappelons aussi qu'une formation de voies fixe, contrairement à une formation de voies adaptative, ne tient pas compte des données reçus aux capteurs, les filtres de formation de voies sont construits pour un ensemble de directions de visée désirées fixes.

Dans ce chapitre, nous proposons un algorithme de séparation aveugle de sources audio, où une formation de voies fixe est utilisée comme une étape de prétraitement à l'algorithme de séparation de sources qui peut être dans notre cas l'ACI ou la minimisation de la norme l_1 (ou dans le cas général, n'importe quel autre algorithme de séparation aveugle de sources audio). Notre formation de voies est construite en utilisant les HRTF comme vecteurs directionnels comme nous l'avons présenté à la section 3.5. Dans ce qui suit, nous présenterons d'abord le principe de la séparation de sources à deux étapes avec une formation de voies comme étape de prétraitement, ensuite nous détaillerons les différentes configurations de ce prétraitement.

5.1 Séparation de sources à deux étapes : principe

Nous rappelons que notre objectif est de trouver, à chaque fréquence, une matrice de séparation $\mathbf{F}(f)$ qui conduira à l'estimation des sources originales dans le domaine temps-fréquence :

$$\mathbf{Y}(f,k) = \mathbf{F}(f)\,\mathbf{X}(f,k) \tag{5.1}$$

avec $\mathbf{Y}(f,k) = [Y_1(f,k),\ldots,Y_N(f,k)]^H$ (*cf.* section 2.1.1). La matrice de séparation $\mathbf{F}(f)$ est estimée en utilisant un algorithme de séparation à deux étapes :

1. une étape de formation de voies fixe : nous filtrons les signaux à la sortie des capteurs avec les filtres de formation de voies $\mathbf{B}(f)$ préalablement estimés hors-ligne, les signaux de sorties sont $\mathbf{Z}(f,k) = \mathbf{B}(f)\,\mathbf{X}(f,k)$;
2. une étape de séparation de sources : nous appliquons un algorithme de séparation de sources (minimisation de la norme l_1 ou l'ACI) à la sortie de la formation de voies $\mathbf{Z}(f,k)$, les signaux de sorties sont les sources estimées $\mathbf{Y}(f,k) = \mathbf{W}(f)\,\mathbf{Z}(f,k)$.

La matrice de séparation finale $\mathbf{F}(f)$ s'écrit comme la combinaison des résultats de ces deux étapes (*cf.* figure 5.1 et algorithme 5.1) :

$$\mathbf{F}(f) = \mathbf{W}(f)\,\mathbf{B}(f) \tag{5.2}$$

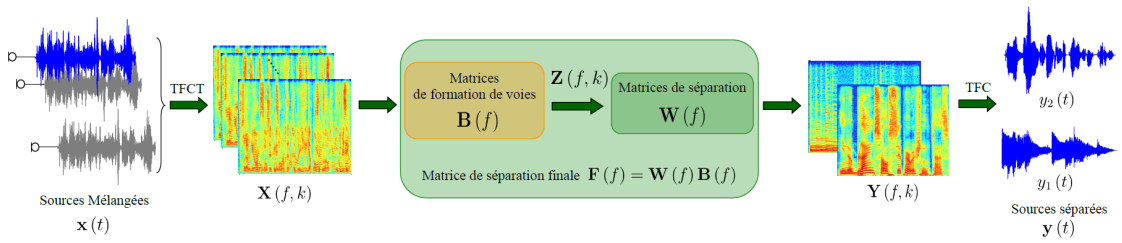


Figure 5.1 – Principe de la séparation de sources avec un prétraitement de formation de voies

Wang *et al.* [89] proposent d'utiliser le prétraitement avec la formation de voies où les directions de visée sont les directions d'arrivées (DOA) des sources. Dans ce cas, les DOA des sources sont supposées être connues *a priori*. Les auteurs ont évalué leur méthode dans un cas déterminé de 2 et 4 sources avec un réseau de capteurs circulaire. Saruwatari *et al.* [72] présentent une méthode combinant analyse en composantes indépendantes et formation de voies : d'abord les auteurs appliquent

Algorithme 5.1 Algorithme général de séparation de sources avec prétraitement de formation de voies

1. **Entrées :**
   (a) Le mélange des sources $\mathbf{x} = [\mathbf{x}(t_1), \ldots, \mathbf{x}(t_T)]$
   (b) Les filtres de formation de voies précalculés $\{\mathbf{B}(f)\}_{1 \leq f \leq \frac{N_f}{2}+1}$
2. $\{\mathbf{X}(f,k)\}_{1 \leq f \leq N_f, 1 \leq k \leq N_T} = \text{TFCT}(\mathbf{x})$
3. Pour chaque fréquence f
   (a) filtrage par filtres de formation de voies : $\mathbf{Z}(f,:) = \mathbf{B}(f)\,\mathbf{X}(f,:)$
   (b) initialiser $\mathbf{W}(f)$: $\mathbf{W}(f) = \mathbf{W}_0(f)$
   (c) initialiser $\mathbf{Y}(f,:)$: $\mathbf{Y}_0(f,:) = \mathbf{W}_0(f)\,\mathbf{Z}(f,:)$
   (d) pour chaque itération t :
   étape de séparation de sources : $[\mathbf{Y}_{t+1}(f,:), \mathbf{W}_{t+1}(f)] = \mathsf{BSS}(\mathbf{Z}(f,:), \mathbf{W}_0)$
4. Résolution du problème de permutation
5. **Sortie :** les sources estimées $\mathbf{y} = \text{TFCT}^{-1}\left(\{\mathbf{Y}(f,k)\}_{1 \leq f \leq N_f, 1 \leq k \leq N_K}\right)$

une ACI dans le domaine fréquentiel et estiment les directions d'arrivées des sources, ensuite ils utilisent les directions d'arrivées estimées pour construire une formation de voies et pour finir, ils intègrent l'ACI et la formation de voies en sélectionnant, à chaque fréquence, la matrice de séparation la plus adaptée.

Nous proposons d'utiliser un prétraitement avec une formation de voies fixe et des directions de visée fixes, indépendamment des directions d'arrivées des sources. Nos filtres de formation de voies $\mathbf{B}(f)$ sont estimés *a priori* en utilisant les HRTF comme expliqué à la section 3.5, pour des directions de visée qui couvrent l'espace utile des directions d'arrivées des sources. Dans le chapitre 9 relatif aux résultats des algorithmes itératifs, nous comparons cette méthode de prétraitement à celle proposée par Wang *et al.* [89] mais dans notre cadre d'étude qui est l'audition des robots. Nous nous intéressons à l'étude des effets de la formation de voies comme un outil de prétraitement, nous n'allons donc pas inclure l'algorithme de Saruwatari *et al.* [72] dans notre évaluation (les auteurs ont utilisé la formation de voies comme méthode de séparation en alternance avec l'ACI).

5.2 Prétraitement avec une formation de voies fixe

Le rôle de la formation de voies est essentiellement de réduire la réverbération et les interférences venant de directions autres que celles qui nous intéressent. Une fois la réverbération réduite, l'équation 2.2 est mieux satisfaite ce qui conduit à une qualité de séparation améliorée. Nous considérons l'ensemble de filtres de formation de voies $\{\mathbf{B}(f)\}_{1\leq f\leq \frac{N_f}{2}+1}$ de dimension $K \times M$, où K est le nombre de lobes désirés, $K \geq N$ et $\mathbf{B}(f) = [\mathbf{b}(f,\theta_1), \ldots, \mathbf{b}(f,\theta_K)]^T$, $\theta_1, \ldots, \theta_K$ étant des directions de visée. Ces filtres de formation de voies sont calculés pour des directions de visée couvrant l'espace utile de mesure. Par exemple, si nous voulons couvrir le demi-plan devant le réseau de capteurs, nous choisissons des formations de voies allant de θ_1=-90° à θ_K=90° avec un angle inter-lobes de 5°, ce qui résulte en $K = 37$ filtres de formation de voies.

Nous rappelons que ces filtres sont calculés à l'avance, avant le début du traitement, et sont utilisés dans le prétraitement par formation de voies comme nous le détaillerons dans les sous-sections suivantes. Dans la suite, nous présenterons différentes configurations du prétraitement par formation de voies.

5.2.1 Formation de voies vers les directions d'arrivées

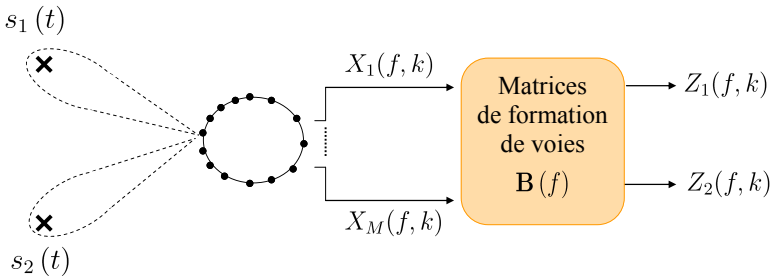


Figure 5.2 – Formation de voies vers les directions d'arrivées

Si les directions d'arrivées des sources $\mathbf{\Phi}^{\text{DOA}} = [\varphi_1, \ldots, \varphi_N]$ sont connues *a priori,* principalement par une méthode de localisation de sources, les filtres de formation de voies sont sélectionnés en utilisant cette information spatiale de localisation des sources. Dans ce cas, les directions de visée désirées sont les directions d'arrivées des sources (*cf.* figure 5.2) et nous sélectionnons les HRTFs correspondants pour construire le vecteur directionnel désiré $\mathbf{d}(f,\theta)$. Dans la pratique, ceci revient à sélectionner, parmi les filtres de formation de voies que nous avons déjà calculés, ceux qui visent vers ces N directions d'arrivées (cf. algorithme 5.2). Le nombre de

Algorithme 5.2 Formation de voies vers les directions d'arrivées BF_DOA

1. **Entrées :**
   (a) Les signaux mélangés $\{\mathbf{X}(f,k)\}_{1\leq f\leq N_f, 1\leq k\leq N_T}$
   (b) Les filtres de formation de voies $\{\mathbf{B}(f)\}_{1\leq f\leq \frac{N_f}{2}+1} = [\mathbf{b}(f,\theta_1),\ldots,\mathbf{b}(f,\theta_K)]^T$
   (c) Les directions d'arrivées $\mathbf{\Phi}^{\mathrm{DOA}} = [\varphi_1,\ldots,\varphi_N]$
2. pour chaque fréquence f
   (a) sélection des filtres visant les directions d'arrivée : $\tilde{\mathbf{B}}(f) = [\mathbf{b}(f,\varphi_1),\ldots,\mathbf{b}(f,\varphi_N)]^T$
   (b) filtrage par formation de voies : $\mathbf{Z}(f,:) = \tilde{\mathbf{B}}(f)\,\mathbf{X}(f,:)$
3. **Sortie :** Signaux filtrés $\mathbf{Z}(f,:)$ de dimension $N \times N_T$

lobes formé est donc égal au nombre de sources N et nous nous trouvons après le filtrage par formation de voies dans un cas de séparation déterminé avec N sources et K mélanges, où $K = N$.

Former des voies vers les directions d'arrivées des sources est une méthode idéale pour comparer avec nos résultats [89]. En effet, pour cette méthode, nous ne nous intéresserons pas à la localisation de sources. Dans [89], où la formation de voies vers les directions d'arrivées a été proposée pour un réseau de capteurs circulaire, les auteurs ont supposé que les directions d'arrivée sont connues *a priori*.

L'algorithme 5.2 présente les étapes de ce filtrage par formation de voies vers les directions d'arrivée que nous appelons BF_DOA :
$\mathbf{Z}(f,:) =$BF_DOA$\big(\mathbf{X}(f,:),\mathbf{B}(f),\mathbf{\Phi}^{\mathrm{DOA}}\big)$.

5.2.2 Formation de voies vers des directions de visée fixes

Estimer les directions d'arrivée des sources pour construire la formation de voies vers ces directions prend un certain temps de traitement et n'est pas toujours précis dans les environnements réverbérants. Nous proposons de construire K lobes fixes, avec des directions de visée arbitraires, choisies de telles sortes qu'elles couvrent l'espace utile des directions d'arrivées des sources. (*cf.* figure 5.3). Nous utilisons les sorties des K filtres de formation de voies $\mathbf{B}(f) = [\mathbf{b}(f,\theta_1),\ldots,\mathbf{b}(f,\theta_K)]^T$ directement dans l'algorithme de séparation. Dans ce cas, nous continuons à avoir un problème de séparation sur-déterminé avec N sources et K mélanges, $K > N$.

L'algorithme 5.3 présente les étapes de ce filtrage par formation de voies fixe que

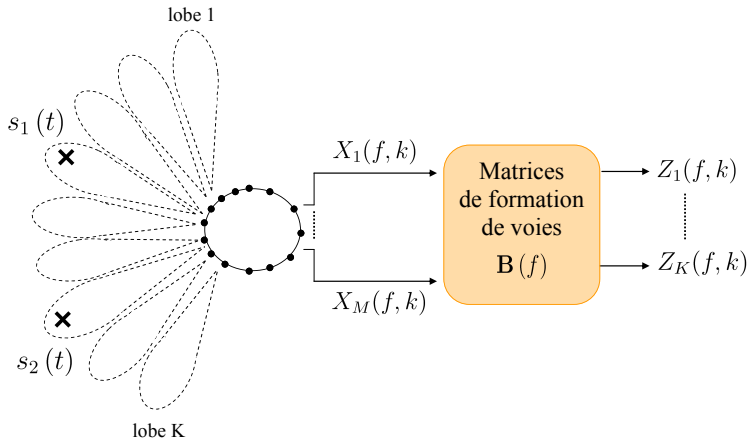


Figure 5.3 – Formation de voies vers des directions de visée fixes

Algorithme 5.3 Formation de voies fixe BF_fixed

1. **Entrées :**
   (a) Les signaux mélangés $\{\mathbf{X}(f,k)\}_{1\leq f\leq \frac{N_f}{2}+1,1\leq k\leq N_T}$
   (b) Les filtres de formation de voies $\{\mathbf{B}(f)\}_{1\leq f\leq \frac{N_f}{2}+1} = [\mathbf{b}(f,\theta_1),\ldots,\mathbf{b}(f,\theta_K)]^T$
2. pour chaque fréquence f
   (a) filtrage par formation de voies : $\mathbf{Z}(f,:) = \mathbf{B}(f)\,\mathbf{X}(f,:)$
3. **Sortie :** Signaux filtrés $\mathbf{Z}(f,:)$ de dimension $K \times N_T$

nous appelons BF_fixed : $\mathbf{Z}(f,:)$ =BF_fixed($\mathbf{X}(f,:), \mathbf{B}(f)$).

5.2.3 Formation de voies vers des directions de visée fixes avec sélection de lobes

Supposons que nous avons effectué la formation de voies vers K directions de visée fixes comme nous l'avons présenté dans la section précédente. Dans cette variante, nous n'utiliserons pas toutes les sorties de cette formation de voies. Si nous supposons que le nombre de sources N est connu *a priori*, nous sélectionnons les N filtres de formation de voies ayant les sorties qui contiennent le maximum d'énergie et qui correspondent donc aux lobes les plus proches des sources (nous supposons que les énergies des sources sont à peu près du même niveau). Après filtrage par les filtres de formation de voies sélectionnés, nous nous trouvons dans un cas de

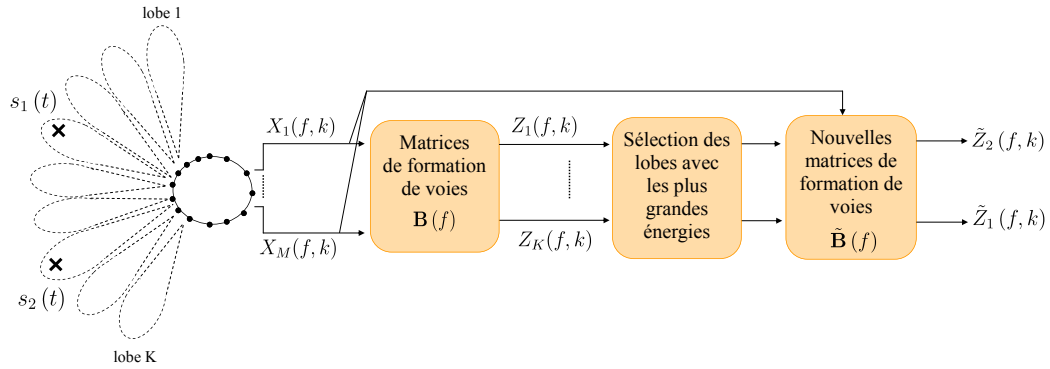


FIGURE 5.4 – Formation de voies fixes avec sélection de lobes contenant les plus grandes énergies

séparation déterminé avec N sources et K mélanges avec $K = N$ (*cf.* figure 5.4).

La sélection de lobes pour la formation de voies nous permet de travailler avec un nombre de mélanges inférieur aux nombre de mélanges de départ et seulement avec les signaux contenant le plus d'information sur les sources utiles.

L'algorithme 5.4 présente les étapes de ce filtrage par formation de voies fixe avec sélection de lobes que nous appelons BF_fixed_BS :
$\mathbf{Z}(f,:) =$BF_fixed_BS$(\mathbf{X}(f,:), \mathbf{B}(f), N)$

5.3 Estimation du nombre de sources et des directions d'arrivées

Dans un scénario de séparation de sources complètement aveugle, le nombre de sources et leurs directions d'arrivée sont inconnus et nous sommes donc amenés à les estimer. La formation de voies vers des directions de visée fixes avec la sélection de lobes peut être dérivée pour l'estimation du nombre de sources ainsi qu'une estimation approximée des directions d'arrivée.

Après le filtrage de formation de voies, le signal est filtré vers K directions de visée choisies *a priori*. Les lobes les plus proches des sources capturent la plus grande partie de leurs énergies. A partir de cette observation, nous proposons d'estimer le nombre de sources en sélectionnant les lobes contenant les plus grandes énergies. Dans ce cas, nous ne connaissons pas le nombre de sources N et nous fixons un nombre maximal N_{max} de sources que peut contenir le mélange. La procédure est la suivante :

1. A chaque bin fréquentiel f, après le filtrage par formation de voies, nous sé-

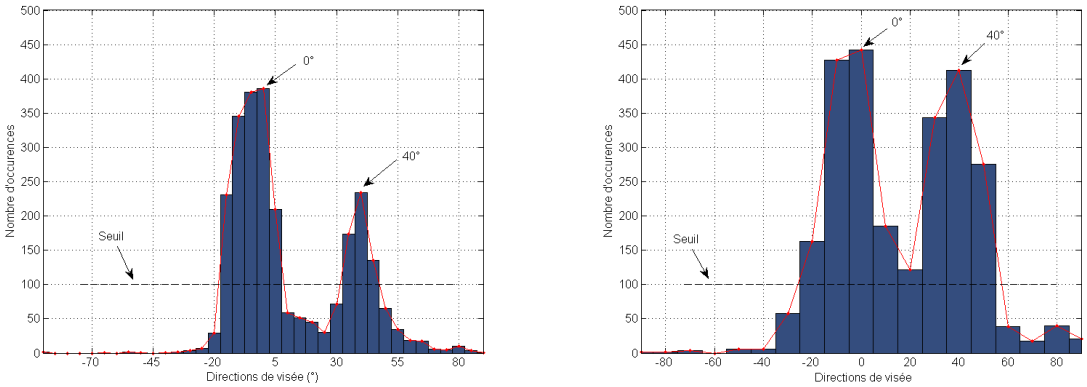
lectionnons N_{max} directions de visée qui correspondent aux N_{max} lobes qui donnent les plus grandes énergies.

2. Nous construisons pour toutes les directions de visée sélectionnées un histogramme qui correspond à leur nombre d'occurrence comme le montre la figure 5.5.

3. Après le seuillage adéquat, nous sélectionnons les pics de cet histogramme. Les filtres qui correspondent à ces directions de visée sont nos filtres de formation de voies sélectionnés $\tilde{\mathbf{B}}(f)$, le nombre de pics correspond à une estimation du nombre de sources $\tilde{N}$ et les directions de visée correspondantes nous donnent une approximation des directions d'arrivée des sources.

Après ce traitement, les signaux filtrés sont $\tilde{\mathbf{Z}}(f,k) = \tilde{\mathbf{B}}(f)\,\mathbf{X}(f,k)$ et le nombre de mélanges correspond au nombre des sources estimées $\tilde{N}$. L'estimation des directions d'arrivée est d'autant plus fine que le nombre de lobes qui couvre l'espace utile est grand : des lobes séparés de 30° donnent une résolution d'estimation de 30° alors que des lobes séparés de 5° donnent une résolution d'estimation des directions d'arrivée de 5° (*cf.* figure 5.5).

L'algorithme 5.5 présente les étapes de ce filtrage par formation de voies fixe avec estimation du nombre de sources que nous appelons BF_fixed_NbSrEstim :
$\left[\mathbf{Z}(f,:),\, \tilde{N},\, \tilde{\mathbf{\Phi}}^{\mathrm{DOA}}\right]$ =BF_fixed_NbSrEstim$(\mathbf{X}(f,:),\, \mathbf{B}(f),\, N_{max},\, K_{\mathrm{thresh}})$



(a) Angle inter-lobes = 5° (b) Angle inter-lobes = 10°

Figure 5.5 – Estimation du nombre de sources en utilisant la formation de voies fixe

Conclusion

Dans ce chapitre, nous avons présenté notre classe d'algorithme de séparation à deux étapes : une étape de prétraitement avec formation de voies fixe utilisant les HRTF et une étape de séparations de sources. Nous avons présenté différentes configurations de cet algorithme en se basant sur différents traitements de la formation de voies fixe dont un résumé est présenté dans le tableau 5.1.

Les principaux avantages de notre méthode sont le temps de calcul réduit par rapport aux algorithmes basés sur la formation de voies adaptative, son amélioration de la qualité de séparation et sa convergence relativement rapide (voir chapitre 9). Ses inconvénients consistent en le manque de preuves ou d'analyses théoriques et en une localisation de sources approximative dont la résolution est égale à l'angle entre deux lobes de la formation de voies fixe.

Méthodes	Sélection de lobes	Entrées	Sorties
BF_DOA	non	Les mélanges $\mathbf{X}(f,:)$ de dimension $M \times N_T$, Les filtres de formation de voies $\mathbf{B}(f)$ de dimension $K \times M$ Les directions d'arrivées $\mathbf{\Phi}^{\text{DOA}}$	$\mathbf{Z}(f,:)$ de dimension $N \times N_T$
BF_fixed	non	Les mélanges $\mathbf{X}(f,:)$ de dimension $M \times N_T$ Les filtres de formation de voies $\mathbf{B}(f)$ de dimension $K \times M$	$\mathbf{Z}(f,:)$de dimension $K \times N_T$
BF_fixed_BS	oui	Les mélanges $\mathbf{X}(f,:)$ de dimension $M \times N_T$ Les filtres de formation de voies $\mathbf{B}(f)$ de dimension $K \times M$ Le nombre de sources N	$\mathbf{Z}(f,:)$de dimension $N \times N_T$
BF_fixed_NbSrEstim	oui	Les mélanges $\mathbf{X}(f,:)$ de dimension $M \times N_T$ Les filtres de formation de voies $\mathbf{B}(f)$ de dimension $K \times M$ Le nombre de lobes maximal à sélectionner N_{max} Le seuil K_{thresh}	$\mathbf{Z}(f,:)$de dimension $\tilde{N} \times N_T$ Le nombre de source estimé $\tilde{N}$ Les directions d'arrivées estimées $\tilde{\mathbf{\Phi}}^{\text{DOA}}$

Table 5.1 – Résumé des méthodes du prétraiement par formation de voies, où $1 \leq f \leq \frac{N_f}{2} + 1$

Algorithme 5.4 Formation de voies fixe avec sélection de lobes, le nombre de source est supposé connu *a priori,* BF_fixed_BS

1. **Entrées :**
   (a) Les signaux mélangés $\{\mathbf{X}(f,k)\}_{1\leq f\leq N_f, 1\leq k\leq N_T}$
   (b) Les filtres de formation de voies $\{\mathbf{B}(f)\}_{1\leq f\leq \frac{N_f}{2}+1} = [\mathbf{b}(f,\theta_1),\ldots,\mathbf{b}(f,\theta_K)]^T$
   (c) Le nombre de sources N
2. SelectedBeams = $\emptyset$
3. pour chaque fréquence f :
   (a) Filtrer les signaux $\mathbf{X}(f,:)$ par les K filtres de formations de voies : $\mathbf{Z}(f,:) = \mathbf{B}(f)\mathbf{X}(f,:)$, $\mathbf{Z}(f,:) = [\mathbf{z}_1(f,:),\ldots,\mathbf{z}_K(f,:)]^T$
   (b) Calculer l'énergie de sortie $\mathbf{E}(f) = [e_1(f),\ldots,e_K(f)]$ des signaux filtrés, avec $e_i(f) = \frac{1}{N_T}\sum_{k=1}^{N_T}|\mathbf{z}_i(f,k)|^2$
   (c) Ordonner $\mathbf{E}(f)$ dans l'ordre décroissant, mettre les valeurs ordonnés dans la variable "Beams", Beams = $sort(\mathbf{E}(f))$
   (d) Sélectionner les N plus grandes énergies, les indices correspondant sont sauvegardé dans B
   (e) SelectedBeams = SelectedBeams $\cup$ B : garder les nouveaux indices avec ceux déjà sélectionnés
4. Calculer la fréquence d'apparition de chaque lobe (de chaque direction de visée) dans "SelectedBeams", les sauvegarder dans I
5. Sélectionner les N lobes ayant les plus grandes occurrences, sauvegarder les filtres correspondant dans $\left\{\tilde{\mathbf{B}}(f)\right\}_{1\leq f\leq \frac{N_f}{2}+1}$
6. **Sortie :** Signaux filtrés $\mathbf{Z}(f,:)$ de dimension $N \times N_T$

Algorithme 5.5 Formation de voies fixe avec sélection de lobes et estimation du nombre de sources, BF_fixed_NbSrEstim

1. **Entrées :**
   (a) Les signaux mélangés dans le domaine temps-fréquence $\{\mathbf{X}(f,k)\}_{1\leq f\leq \frac{N_f}{2}+1,1\leq k\leq N_T}$
   (b) Les filtres de formation de voies pré-calculés $\{\mathbf{B}(f)\}_{1\leq f\leq \frac{N_f}{2}+1} = [\mathbf{b}(f,\theta_1),\ldots,\mathbf{b}(f,\theta_K)]^T$
   (c) Le nombre de sources maximum N_{max}
   (d) Le seuil K_{thresh}
2. SelectedBeams = $\emptyset$
3. pour chaque fréquence f :
   (a) Filtrer les signaux $\mathbf{X}(f,:)$ par les K filtres de formations de voies : $\mathbf{Z}(f,:) = \mathbf{B}(f)\mathbf{X}(f,:)$, $\mathbf{Z}(f,:) = [\mathbf{z}_1(f,:),\ldots,\mathbf{z}_K(f,:)]^T$
   (b) Calculer l'énergie de sortie $\mathbf{E}(f) = [e_1(f),\ldots,e_K(f)]$ des signaux filtrés, avec $e_i(f) = \frac{1}{N_T}\sum_{k=1}^{N_T}|\mathbf{z}_i(f,k)|^2$
   (c) Ordonner $\mathbf{E}(f)$ dans l'ordre décroissant, mettre les valeurs ordonnés dans la variable "Beams", Beams = $sort(\mathbf{E}(f))$
   (d) Sélectionner les N_{max} plus grandes énergies, les indices correspondants sont sauvegardé dans B
   (e) SelectedBeams = SelectedBeams $\cup$ B
4. Calculer la fréquence d'apparition de chaque lobe (de chaque direction de visée) dans "SelectedBeams", les sauvegarder dans I
5. Seuiller I selon le seuil K_{thresh} et détecter les pics qui s'y trouvent I_{max}
6. $\tilde{N}$ est le nombre de pics détecté, $\tilde{\mathbf{\Phi}}^{\text{DOA}}$sont les directions de visée correspondantes à ces pics
7. **Sorties :**
   (a) Signaux filtrés $\mathbf{Z}(f,:)$ de dimension $\tilde{N}\times N_T$
   (b) Le nombre de sources estimé $\tilde{N}$
   (c) Les directions d'arrivées estimées $\tilde{\mathbf{\Phi}}^{\text{DOA}}$

Chapitre 6

Séparation adaptative de sources audio avec prétraitement de formation de voies

Introduction

Dans un cas de séparation *aveugle* de sources audio dans un environnement *réel*, nous ne connaissons pas les conditions dans lesquelles les signaux sources arrivent aux capteurs. En effet, le nombre de sources, leurs positions, le bruit environnant et le taux de réverbération de la pièce sont inconnus. De plus, ces conditions d'*écoute* ne sont pas fixes au cours du temps mais peuvent changer dynamiquement. Par exemple dans une conversation entre deux locuteurs, nous pouvons avoir à un instant t zéro, une ou deux sources sonores. Pour pouvoir s'adapter aux changements de l'environnement dans lequel il évolue, le robot doit procéder à une séparation adaptative et en temps réel des sources audio. Nous nous limitons dans cette thèse à l'aspect adaptatif des algorithmes de séparation de sources.

Dans ce chapitre, nous proposons une version adaptative de notre algorithme de séparation de sources avec un prétraitement de formation de voies présenté dans le chapitre précédent. Nous supposons dans un premier temps que le nombre de sources est fixe au cours du temps, puis nous passerons à un cas plus réaliste dans lequel le nombre de sources est variable au cours du temps.

6.1 Schéma d'adaptation

Dans un algorithme de séparation adaptative de sources audio, la séparation des sources s'effectue sur un certain nombre d'échantillons des signaux reçus qu'on appellera *fenêtre d'analyse longue*. La longueur de cette fenêtre correspond à la taille du buffer utilisé dans le DSP. Une fois les signaux séparés sur une fenêtre d'analyse longue, l'algorithme passe à la fenêtre suivante et procède au même traitement. Dans cette section, nous définissons dans un premier temps les différentes fenêtres d'analyse utilisées dans notre algorithme adaptatif, ensuite nous présentons le principe d'adaptation de cet algorithme de séparation avec prétraitement de formation de voies, et nous finissons sur les problèmes rencontrés dans la reconstruction des signaux audio séparés.

6.1.1 Fenêtres d'analyse

Fenêtre d'analyse spectrale Les signaux audio sont souvent analysés sur des fenêtres d'analyse temporelle assez courte, de l'ordre de la durée de stationnarité du signal. En effet, les propriétés fréquentielles et temporelles de ce genre de signaux varient d'une manière significative et dépendante de la source étudiée. Un compromis entre résolution temporelle qui nécessite une fenêtre d'analyse courte et résolution spectrale qui nécessite une fenêtre d'analyse longue est nécessaire. Nous adoptons une fenêtre d'analyse de type Hamming avec une longueur de $T = 1024$ échantillons, ce qui correspond à une durée de 64 ms pour une fréquence d'échantillonnage de 16 KHz, avec un pas d'avancement de 32 ms. Ces valeurs correspondent à celles choisies pour effectuer l'analyse spectrale à l'aide de la transformée de Fourier à court terme (TFCT).

Fenêtre d'analyse longue L'algorithme de séparation de sources a besoin d'un certain nombre de fenêtres d'analyse pour pouvoir procéder correctement à la séparation. Nous définissons une fenêtre d'analyse temporelle plus longue que celle utilisé pour l'analyse spectrale des signaux, $T_L = 16T = 1s$. Cette fenêtre définit la longueur du signal temporel qui sera transmis à l'algorithme adaptatif de séparation de sources. C'est une fenêtre d'analyse glissante dont le pas d'avancement est égal à la moitié de la longueur de la fenêtre d'analyse spectrale, c'est à dire 512 échantillons (cf. figure 6.1).

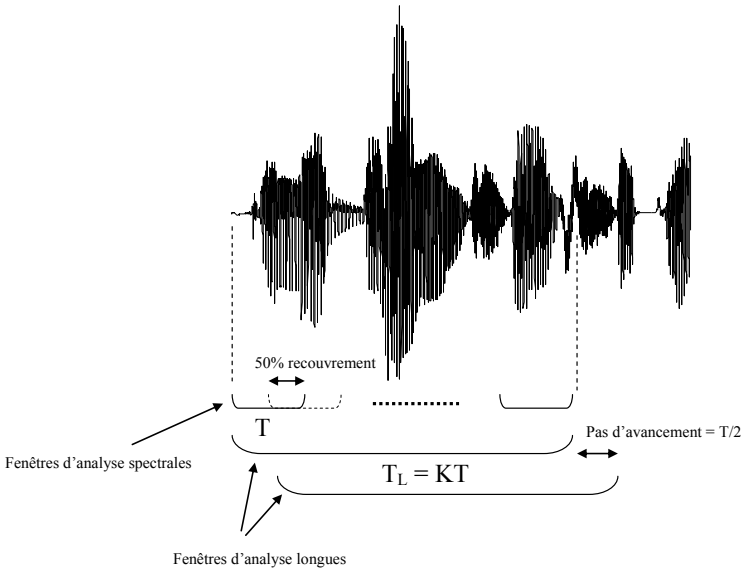


FIGURE 6.1 – Les fenêtres d'analyse utilisées dans la séparation adaptative des sources audio

6.1.2 Principe d'adaptation

Première fenêtre d'analyse

Supposons qu'une source sonore vient de s'activer dans l'environnement du robot, le traitement appliqué à la première fenêtre longue (la première seconde du signal) des signaux est comme suit :

1. nous appliquons une transformée de Fourier sur les $T = 1024$ premiers échantillons reçus et nous sauvegardons le résultat ;
2. nous concaténons les $\frac{T}{2}$ nouveaux échantillons temporel du signal reçu avec les $\frac{T}{2}$ échantillons précédents et nous exécutons une transformée de Fourier ; les étapes 1 et 2 reviennent à faire une transformée de Fourier à court terme sur le signal temporel dans un cas de séparation hors-ligne, avec une fenêtre d'analyse fréquentielle de longueur $T = 1024$ échantillons et un recouvrement de 50% ;
3. dès qu'un nombre d'échantillons égal à celui de la fenêtre d'analyse longue est reçu, c'est à dire $T_L = 16T = 1s$, le processus de séparation de sources commence :
   (a) la première étape consiste en un filtrage par formation de voies du signal en utilisant les filtres de formation de voies calculés hors-ligne et

l'une des méthodes suivantes : BF_DOA, BF_fixed, BF_fixed_BS ou BF_fixed_NbSrEstim ;

(b) la deuxième étape consiste en la séparation aveugle des signaux filtrés en utilisant la version adaptative d'un des algorithmes de séparation aveugle de source audio BSS présenté dans le chapitre 4 ; puisque c'est la première seconde du signal que nous analysons, nous avons besoin d'initialiser les matrices de séparations ;

4. nous reconstituons le signal dans le domaine temporel avec une transformée de Fourier à court terme inverse.

$\mathbf{t}^{ième}$ fenêtre d'analyse

Supposons que nous avons reçu et traité $t-1$ fenêtres longues du signal temporel mélangé. A ce stade du processus, nous disposons des matrices de séparations dans le domaine fréquentiel $\mathbf{W}_{t-1}(f)$. Voici le traitement appliqué à la $t^{ième}$ fenêtre longue :

1. nous concaténons les $\frac{T}{2}$ nouveaux échantillons aux $\frac{T}{2}$ derniers échantillons de la $t-1^{ième}$ fenêtre et nous procédons à une FFT ;
2. nous concaténons cette FFT aux 15 dernières FFT de la $t-1^{ième}$ fenêtre pour former la $t^{ième}$ fenêtre d'analyse de longueur $T_L = 16T = 1s$;
3. sur cette $t^{ième}$ fenêtre, nous procédons à la séparation de sources :

   (a) la première étape consiste en un filtrage par formation de voies du signal en utilisant les filtres de formation de voies calculés hors-ligne ;

   (b) la deuxième étape consiste en la séparation aveugle des signaux filtrés en utilisant la version adaptative d'un des algorithmes de séparation aveugle de source audio présenté dans le chapitre 4 en l'initialisant avec les matrices de séparation estimées à la fenêtre précédente $\mathbf{W}_{t-1}(f)$;

4. nous reconstruisons le signal dans le domaine temporel avec une transformée de Fourier à court terme inverse.

Les étapes 1, 2, 3.b et 4 représente la version adaptative des algorithmes de séparation aveugle de sources audio présentés dans le chapitre 4, c'est-à-dire les algorithmes de séparation sans prétraitement par formation de voies. La figure 6.2 résume ce schéma d'adaptation.

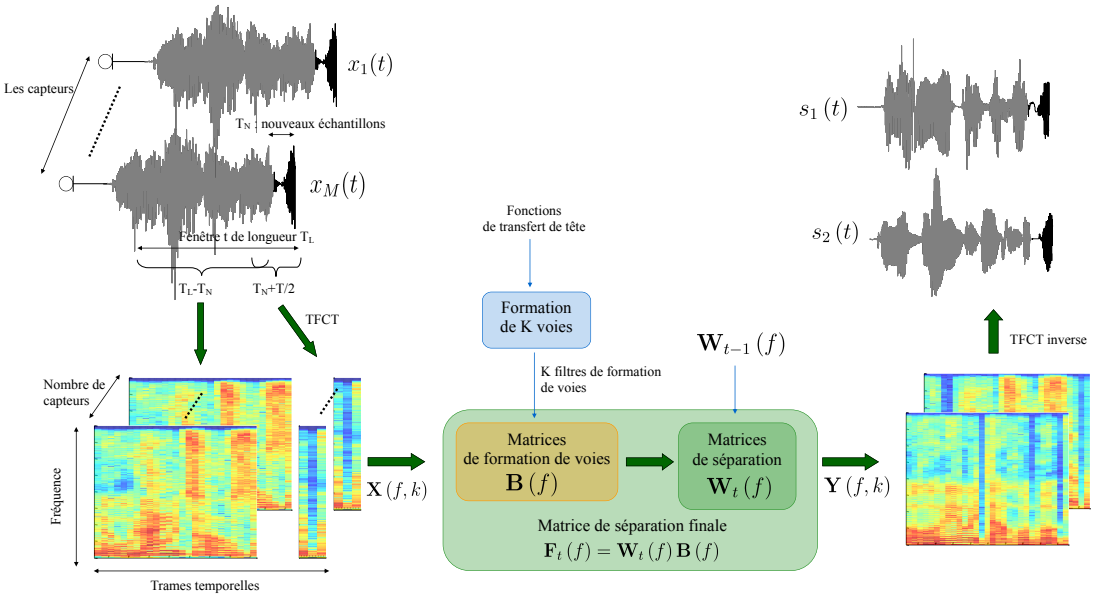


FIGURE 6.2 – Schéma d'adaptation pour l'algorithme de séparation de sources avec un prétraitement par formation de voies

6.1.3 Problèmes de permutation et d'échelle dans le domaine temporel

Dans une séparation adaptative des sources, les signaux sont estimés indépendamment sur chaque trame d'analyse longue. L'ordre des signaux sources estimés ainsi que leurs échelles peuvent être différent d'une fenêtre d'analyse longue à la suivante. Il faut donc penser à corriger l'ordre et l'échelle des sources estimés en exploitant leurs corrélations.

6.2 Algorithme de séparation sans estimation du nombre de sources

Si nous choisissons de ne pas estimer le nombre de sources, deux cas de figure se présentent :

1. le nombre de sources est supposé connu *a priori* ;
2. le nombre de sources est inconnu.

La séparation de sources dans un scénario réel implique que le nombre de sources est variable et inconnu. Si nous choisissons de ne pas procéder à une estimation de ce nombre de sources à chaque fenêtre d'analyse, une solution consiste à introduire dans

l'algorithme de séparation un nombre de source à séparer fixe N_{hyp}, ce nombre reste le même tout au long de l'algorithme, quel que soit le nombre de source actif dans une fenêtre d'analyse donnée et pour toutes les fenêtres d'analyse. Ce nombre de sources est choisi arbitrairement, de telle sorte que le nombre maximal de sources dans l'environnement du réseau de capteurs ne le dépasse pas : $N_t \leq N_{hyp}, \forall t$. Ceci dépend évidemment de l'environnement dans lequel nous voulons effectuer la séparation de sources. Dans notre cas, c'est une séparation de sources dans un appartement dans lequel vit une personne âgée, N_{hyp} ne devrait donc pas être un grand nombre.

Dans le cas où le nombre de sources est connu *a priori* ou dans celui où ce nombre est inconnu et fixe, nous utilisons à l'étape de prétraitement une formation de voies fixe avec sélection de lobes BF_fixed_BS. L'algorithme 6.1 présente les étapes de la séparation de sources adaptative sans estimation du nombre de sources, nous appelons cette méthode de séparation BF_fixed_BS+BSS.

6.3 Algorithme de séparation avec estimation du nombre de sources

Dans ce cas, le nombre de sources à séparer est estimé à chaque fenêtre d'analyse. Le traitement à adopter à la fenêtre d'analyse t dépend du nombre de sources $\tilde{N}_t$ estimé à cette fenêtre et du nombre de sources $\tilde{N}_{t-1}$ estimé à la fenêtre d'analyse précédente (*cf.* algorithme 6.2). En effet, l'algorithme de séparation de sources BSS utilise à la fenêtre t la matrice de séparation $\mathbf{W}_{t-1}(f)$ estimée à la fenêtre d'analyse précédente afin de s'initialiser, cette matrice est carrée de dimension $\tilde{N}_{t-1} \times \tilde{N}_{t-1}$. Si à la fenêtre t une source s'est éteinte ou s'est activée, $\tilde{N}_t$ et $\tilde{N}_{t-1}$ deviennent différents et nous ne pouvons plus initialiser l'algorithme de séparation avec $\mathbf{W}_{t-1}(f)$ à cause de sa dimension incompatible avec celle de $\mathbf{Z}_t(f,k)$.

Selon le nombre estimé de sources à la fenêtre t et à la fenêtre $t-1$, trois cas de figure se présentent :

1. $\tilde{N}_t = \tilde{N}_{t-1}$: aucun traitement particulier n'est nécessaire ;
2. $\tilde{N}_t > \tilde{N}_{t-1}$: une ou plusieurs sources se sont activées à la fenêtre t, il faut modifier la matrice de séparation $\mathbf{W}_{t-1}(f)$ pour qu'elle tienne compte de ces nouvelles sources ;
3. $\tilde{N}_t < \tilde{N}_{t-1}$: une ou plusieurs sources se sont éteintes à la fenêtre t, il faut retirer les lignes et les colonnes de $\mathbf{W}_{t-1}(f)$ correspondant à ces sources.

La modification de la matrice de séparation $\mathbf{W}_{t-1}(f)$ en fonction du nombre de

Algorithme 6.1 Séparation adaptative de sources avec prétraitement de formation de voies : le nombre de sources est connu *a priori* ou inconnu et fixé arbitrairement, BF_fixed_BS+BSS

1. **Entrées :**
   (a) Les filtres de formation de voies $\{\mathbf{B}(f)\}_{1\leq f\leq \frac{N_f}{2}+1} = [\mathbf{b}(f,\theta_1),\ldots,\mathbf{b}(f,\theta_K)]^T$
   (b) Le nombre de sources fixé N_{hyp} ou réel N_t, $\forall t$
2. $t = 0$:
   (a) **Entrée :** Les signaux mélangés à la $1^{ère}$ fenêtre $\{\mathbf{X}_0(f,k)\}_{1\leq f\leq N_f, 1\leq k\leq N_T}$
   (b) Prétraitement par formation de voies : $\mathbf{Z}_0(f,:) =$BF_fixed_BS$(\mathbf{X}_0(f,:),\mathbf{B}(f), N_t$ ou $N_{hyp})$
   (c) Initialiser $\mathbf{W}(f)$: $\mathbf{W}(f) = \mathbf{W}_0(f)$
   (d) Etape de séparation de sources : $[\mathbf{Y}_1(f,:), \mathbf{W}_1(f)] = \mathrm{BSS}(\mathbf{Z}_0(f,:), \mathbf{W}_0)$
3. $t \geq 1$
   (a) **Entrées :**
      i. Les signaux mélangés à la $t^{ième}$ fenêtre $\{\mathbf{X}_t(f,k)\}_{1\leq f\leq N_f, 1\leq k\leq N_T}$
      ii. Les matrices de séparations de la fenêtre précédente $\mathbf{W}_t(f)$
   (b) Prétraitement par formation de voies : $\mathbf{Z}_t(f,:) =$BF_fixed_BS$(\mathbf{X}_t(f,k),\mathbf{B}(f), N$ ou $N_{hyp})$
   (c) Etape de séparation de sources : $[\mathbf{Y}_{t+1}(f,:), \mathbf{W}_{t+1}(f)] =BSS(\mathbf{Z}_t(f,:), \mathbf{W}_t(f))$
4. **Sortie :** Les sources séparées $\{\mathbf{Y}(f,k)\}_{1\leq f\leq N_f, 1\leq k\leq N_T}$

sources estimées à la fenêtre d'analyse courante et précédente se base essentiellement sur l'estimation des directions d'arrivées qui sont apparues ou disparues d'une fenêtre à une autre. Nous rappelons que lors d'un prétraitement par formation de voies avec estimation du nombre de source BF_fixed_NbSrEstim, nous estimons le nombre de sources mais aussi les directions d'arrivée des sources, ces estimations se basent sur la détection des lobes capturant les plus grandes énergies des sources. La résolution des directions d'arrivées estimées est égale à l'angle entre les lobes de la formation de voies fixe. Pour une formation de voies allant de θ_1=-90° à θ_K=90° avec un angle inter-lobes de 5°, la résolution des directions d'arrivées est aussi de 5°. Dans la suite, nous détaillons comment ces matrices de séparation sont modifiées.

6.3.1 Activation d'une ou plusieurs sources

Lorsque, en passant d'une fenêtre d'analyse $t-1$ à une fenêtre d'analyse t, le nombre de source augmente de $d = \tilde{N}_t - \tilde{N}_{t-1}$, la matrice de séparation $\mathbf{W}_{t-1}(f)$ doit être modifiée afin de tenir compte de ces nouvelles sources $\mathbf{s}^{new}(t)$. D'abord, il faut trouver les indices $i_1^{new}, \ldots, i_d^{new}$ des sources qui viennent de s'activer, c'est-à-dire leurs *ordres* par rapport aux sources qui existe depuis la trame $t-1$. Ceci se fait en se basant sur les nouvelles directions d'arrivées détectées à la fenêtre t et leurs positions. Ensuite pour chaque nouvelle source $s_{i_n}^{new}(t)$, $n \in \{1, \ldots, d\}$, nous insérons à la $i_n^{ième}$ ligne et à la $i_n^{ième}$ colonne de $\mathbf{W}_{t-1}(f)$ le vecteur $\left[0, 0, \ldots, 0, \underset{i_n^{ième}}{1}, 0, \ldots, 0\right]$. Nous appelons cette fonction IndSrNew et elle prend comme entrées $\mathbf{W}_{t-1}(f)$, les directions d'arrivée estimées à la fenêtre t et à la fenêtre $t-1$ respectivement $\tilde{\mathbf{\Phi}}_t^{\mathrm{DOA}}$ et $\tilde{\mathbf{\Phi}}_{t-1}^{\mathrm{DOA}}$:$\tilde{\mathbf{W}}_{t-1}(f) = \mathsf{IndSrNew}\left(\mathbf{W}_{t-1}(f), \tilde{\mathbf{\Phi}}_t^{\mathrm{DOA}}, \tilde{\mathbf{\Phi}}_{t-1}^{\mathrm{DOA}}\right)$.

Par exemple, si à la trame $t-1$ l'algorithme détecte deux sources à 0° et 60°, et qu'à la trame t une source située à 30° s'active, la matrice $\tilde{\mathbf{W}}_{t-1}(f)$ s'écrit en fonction de $\mathbf{W}_{t-1}$:

$$\tilde{\mathbf{W}}_{t-1}(f) = \begin{bmatrix} W_{11} & 0 & W_{12} \\ 0 & 1 & 0 \\ W_{21} & 0 & W_{22} \end{bmatrix}$$

6.3.2 Extinction d'une ou plusieurs sources

Lorsque une ou plusieurs sources s'éteignent en passant de la fenêtre d'analyse $t-1$ à la fenêtre d'analyse t, les lignes et les colonnes relatives à ces sources doivent être supprimées de $\mathbf{W}_{t-1}(f)$. Soit $d = \tilde{N}_{t-1} - \tilde{N}_t$ le nombre de sources qui se sont éteintes

et $i_1^{vanish}, \ldots, i_d^{vanish}$ leurs indices. En se basant sur la différence dans les directions d'arrivées des sources estimées dans les deux fenêtres, nous pouvons détecter les sources qui se sont éteintes et supprimer les lignes de $\tilde{\mathbf{W}}_{t-1}(f)$ qui correspondent aux indices détectés. Les colonnes à supprimer correspondent aux mélanges qui se sont éteints.

Nous appelons cette fonction IndSrVanish et elle prend comme entrées $\mathbf{W}_{t-1}(f)$, les directions d'arrivée estimées à la fenêtre t et à la fenêtre $t-1$ respectivement $\tilde{\mathbf{\Phi}}_t^{\mathrm{DOA}}$ et $\tilde{\mathbf{\Phi}}_{t-1}^{\mathrm{DOA}}$, $\mathbf{Y}_{t-1}(f,:)$ et $\mathbf{Y}_t(f,:)$:
$\tilde{\mathbf{W}}_{t-1}(f) = \mathsf{IndSrVanish}\left(\mathbf{W}_{t-1}(f), \tilde{\mathbf{\Phi}}_t^{\mathrm{DOA}}, \tilde{\mathbf{\Phi}}_{t-1}^{\mathrm{DOA}}, \mathbf{Y}_{t-1}(f,:), \mathbf{Y}_t(f,:)\right)$.

Conclusion

Dans ce chapitre, nous avons présenté la version adaptative de notre algorithme de séparation à deux étapes avec un prétraitement de formation de voies fixe, dans un cas où le nombre de sources change au cours de temps. Dans un premier temps, nous supposons que le nombre de sources à séparer n'est pas estimé : soit il est connu *a priori*, soit il est fixé de telle sorte qu'il soit supérieur au nombre réel de sources. Ensuite, nous avons proposé d'estimer le nombre de sources d'une manière adaptative en se basant sur l'étape de formation de voies fixe et la sélection de lobes contenant les plus grandes énergies des sources. L'évaluation de ces algorithmes se fera dans le chapitre 8 après la présentation des bases de données que nous avons développées ainsi que les méthodes de mesures dans le chapitre 7.

Algorithme 6.2 Séparation adaptative de sources avec prétraitement de formation de voies et estimation du nombre de sources à chaque fenêtre, BF_fixed_NbSrEstim+BSS

1. **Entrées :**
   (a) Les filtres de formation de voies $\{\mathbf{B}(f)\}_{1\leq f\leq \frac{N_f}{2}+1} = [\mathbf{b}(f,\theta_1),\ldots,\mathbf{b}(f,\theta_K)]^T$
   (b) Le nombre de lobes sélectionné par fréquence N_{max}
2. $t=0$:
   (a) **Entrée :** Les signaux mélangés à la $1^{ère}$ fenêtre $\{\mathbf{X}_0(f,k)\}_{1\leq f\leq N_f,1\leq k\leq N_T}$
   (b) Prétraitement par formation de voies : $\left[\mathbf{Z}_0(f,:),\tilde{N}_0,\tilde{\mathbf{\Phi}}_0^{\mathrm{DOA}}\right]$ =BF_fixed_NbSrEstim$(\mathbf{X}_0(f,:),\mathbf{B}(f),N_{max},K_{\mathrm{thresh}})$
   (c) Initialiser $\mathbf{W}(f)$: $\mathbf{W}(f)=\mathbf{W}_0(f)$
   (d) Etape de séparation de sources : $[\mathbf{Y}_1(f,:),\mathbf{W}_1(f)]=\mathrm{BSS}(\mathbf{Z}_0(f,:),\mathbf{W}_0)$
3. $t\geq 1$
   (a) **Entrées :**
      i. Les signaux mélangés à la $t^{ième}$ fenêtre $\{\mathbf{X}_t(f,k)\}_{1\leq f\leq N_f,1\leq k\leq N_T}$
      ii. Les matrices de séparations de la fenêtre précédente $\mathbf{W}_t(f)$
      iii. Le nombre de sources estimées à la fenêtre précédente $\tilde{N}_t$
      iv. Les directions d'arrivées estimées à la fenêtre précédente $\tilde{\mathbf{\Phi}}_t^{\mathrm{DOA}}$
   (b) Prétraitement par formation de voies : $\left[\mathbf{Z}_{t+1}(f,:),\tilde{N}_{t+1},\tilde{\mathbf{\Phi}}_{t+1}^{\mathrm{DOA}}\right]$ =BF_fixed_NbSrEstim$(\mathbf{X}_t(f,:),\mathbf{B}(f),N_{max},K_{\mathrm{thresh}})$
   (c) Etape de séparation de sources :
      i. si $\tilde{N}_{t+1}=\tilde{N}_t$, $[\mathbf{Y}_{t+1}(f,:),\mathbf{W}_{t+1}(f)]$ =BSS$(\mathbf{Z}_t(f,:),\mathbf{W}_t(f))$
      ii. si $\tilde{N}_{t+1}>\tilde{N}_t$,
         A. $\tilde{\mathbf{W}}_t(f)=$ IndSrNew $\left(\mathbf{W}_t(f),\tilde{\mathbf{\Phi}}_{t+1}^{\mathrm{DOA}},\tilde{\mathbf{\Phi}}_t^{\mathrm{DOA}}\right)$
         B. $[\mathbf{Y}_{t+1}(f,:),\mathbf{W}_{t+1}(f)]$ =BSS$\left(\mathbf{Z}_t(f,:),\tilde{\mathbf{W}}_t(f)\right)$
      iii. si $\tilde{N}_{t+1}<\tilde{N}_t$,
         A. $\tilde{\mathbf{W}}_t(f)=$ IndSrVanish $\left(\mathbf{W}_t(f),\tilde{\mathbf{\Phi}}_{t+1}^{\mathrm{DOA}},\tilde{\mathbf{\Phi}}_t^{\mathrm{DOA}},\mathbf{Y}_t(f,:),\mathbf{Y}_{t+1}(f,:)\right)$
         B. $[\mathbf{Y}_{t+1}(f,:),\mathbf{W}_{t+1}(f)]$ =BSS$\left(\mathbf{Z}_t(f,:),\tilde{\mathbf{W}}_t(f)\right)$
4. **Sortie :** Les sources séparées $\{\mathbf{Y}(f,k)\}_{1\leq f\leq N_f,1\leq k\leq N_T}$

Quatrième partie

Expériences et résultats

Chapitre 7

Bases de données pour la séparation de sources

Introduction

Dans ce chapitre, nous présentons deux bases de données que nous avons développées dans le cadre de cette thèse pour des utilisations différentes :

1. une base de données de parole mesurée dans des conditions acoustiques différentes, cette base nous sert à évaluer nos algorithmes de séparation de sources et a été mesurée dans deux milieux acoustiques différents ;
2. une base de données de fonctions de transfert de tête (HRTF) utilisée pour construire les filtres de formation de voies.

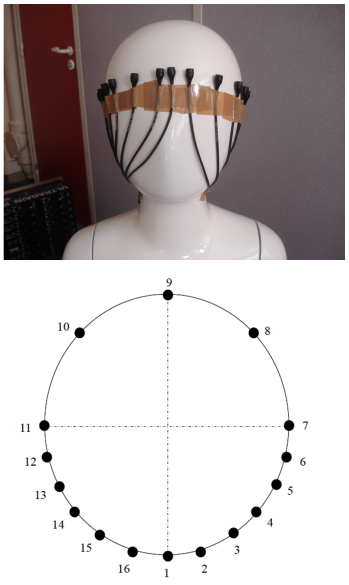
Ces bases de données ont été mesurées avec un réseau de capteurs modélisant le futur robot comme nous le détaillerons dans la première section de ce chapitre. Ces bases de données sont basées sur le calcul des *réponses impulsionnelles* que nous présenterons dans la deuxième section. Le processus d'acquisition sera détaillé dans la dernière section.

7.1 Le réseau de capteurs pour les mesures

Dans le cadre du projet ROMEO, les microphones sont placés sur la tête de l'humanoïde. Dans un premier temps, nous avons modélisé le futur humanoïde par un mannequin de vitrine de taille 1m20 que nous appelons *Theo* (*cf.* figure 7.1a). Nous avons placé 16 capteurs autour de la tête de Théo comme indiqué dans la figure 7.1b. Theo nous a servi à faire les premières mesures en attendant la conception et

la réception du prototype de la tête et torse du robot final Romeo. Pour les mesures, Theo a été mis sur une table tournante, la taille totale du dispositif "Theo + table tournante" est de 1m40 et correspond à la taille de Romeo. Les bases de données enregistrées avec Theo ont pour nom Theo-<nom-de-la-base-de-données>.

(a) Theo, le mannequin de vitrine utilisé pour modéliser le futur robot Romeo



(b) La position des capteurs autour de la tête de Théo (vu de dessus)

Figure 7.1 – Le réseau de capteurs de Theo

7.2 Réponse impulsionnelle acoustique et temps de réverbération

La construction des signaux mélangés et le calcul des HRTF passent par l'estimation des réponses impulsionnelles acoustiques entre différents point de la salle et les capteurs. La *réponse impulsionnelle acoustique* d'un point d'émission vers un capteur caractérise le chemin acoustique entre ces deux points. Elle contient les critères acoustiques de la salle dans laquelle la mesure est faite, typiquement le taux de réverbération de la pièce. Observée dans le domaine temporel, la réponse impulsionnelle acoustique montre les réflexions importantes : le trajet direct, les réflexions précoces et le champ diffus ou réverbéré. L'amplitude maximale de la réponse impulsionnelle correspond à la première onde arrivée, nous pouvons aussi mesurer la durée du trajet direct ainsi que le temps d'arrivée des réflexions précoces. La figure 7.3

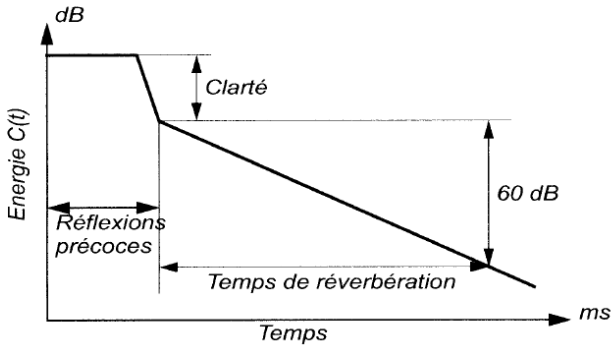


FIGURE 7.2 – Paramètres extraits de la courbe de décroissance de l'énergie (EDC) (cette figure est extraite du manuel de cours d'Yves Grenier [36])

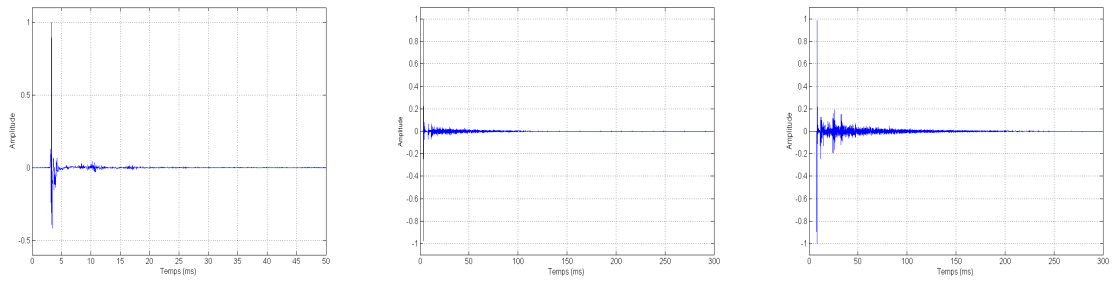
montre différentes réponses impulsionnelles enregistrées dans des conditions acoustiques différentes :

- la figure 7.3a représente une réponse impulsionnelle enregistrée dans une chambre sourde, nous pouvons y distinguer le trajet direct et l'absence quasi-totale de réflexions précoces et de champ diffus ;
- la figure 7.3b représente une réponse impulsionnelle enregistrée dans le studio de Télécom ParisTech ; cette réponse montre que le studio est d'une réverbération modérée, vu l'existence de réflexions précoces avec une amplitude pas très élevée ; le champ diffus quant à lui s'estompe après 100ms ;
- la figure 7.3c est l'estimation de la réponse impulsionnelle de l'Institut de la Vision, nous pouvons distinguer des réflexions précoces d'amplitude plus importante que celle du studio ainsi qu'un champ diffus qui s'estompe plus tardivement ; nous pouvons conclure en regardant ces réponses impulsionnelles que la pièce de l'Institut de la Vision est plus réverbérante que le studio de Télécom ParisTech.

Le calcul du temps de réverbération peut se faire à partir de la *courbe de décroissance de l'énergie* (EDC : Energy Decay Curve). Cette courbe nous permet d'accéder à la durée des réflexions précoces, à la clarté et au temps de réverbération. La courbe de décroissance de l'énergie $C(t)$ est définie comme l'énergie de la réponse impulsionnelle $h(t)$ depuis l'instant t jusqu'à la fin de la réponse, théoriquement $t \to \infty$. Son expression est définie en décibels comme suit :

$$C(t) = 10 \log \sum_{\tau=t}^{\infty} h^2(\tau) \tag{7.1}$$

La figure 7.2 montre l'allure typique de la courbe de décroissance de l'énergie. Le temps d'arrivée du trajet direct est caractérisé par la durée du plateau horizontal. Ensuite, entre le moment où le trajet direct est éliminé de la réponse et les réflexions précoces, la courbe chute brutalement. Ensuite, la courbe EDC décroit régulièrement ce qui correspond à la décroissance exponentielle de la partie de la réponse correspondante au champ diffus ou réverbéré. L'estimation du temps de réverbération se fait à partir de la courbe EDC comme le montre la figure 7.2. Le temps de réverbération est défini comme l'intervalle de temps durant lequel la pression acoustique d'une salle diminue à un millième de sa valeur de régime établi, suite à l'arrêt de la source sonore. Cela représente une diminution du niveau sonore de 60dB et dans ce cas, le temps de réverbération est noté RT_{60}. Dans un environnement réel, il est difficile d'obtenir une décroissance de 60dB du niveau sonore, il est donc plus commun d'utiliser RT_{30} ou RT_{20} qui représentent le temps que prend une source pour décroitre de 30dB ou de 20dB respectivement. La figure 7.4 montre les courbes de décroissance de l'énergie obtenues dans la chambre anéchoïque, dans le studio de Télécom ParisTech et à l'Institut De la Vision.

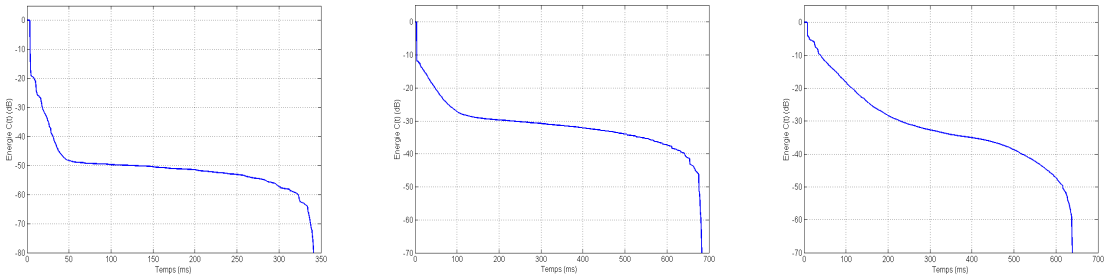


(a) Chambre anéchoïque (b) Studio Télécom ParisTech (c) Institut de la vision

Figure 7.3 – Réponse impulsionnelle entre un point dans une pièce et le microphone 1

7.3 Construction des mélanges convolutifs

Une bonne évaluation des algorithmes de séparation de sources nécessite, à égale importance, des outils d'évaluation des performances de séparation qui donnent une description complète de la *qualité* des signaux séparés (*cf.* chapitre 8 section 8) et une *bonne* base de données d'évaluation. Une base de données d'évaluation des algorithmes de séparation de sources doit être :



(a) Chambre anéchoïque (b) Studio Télécom ParisTech (c) Institut de la vision

FIGURE 7.4 – Courbes de décroissance de l'énergie (EDC) entre un point dans une pièce et le premier microphone

- assez large pour permettre une évaluation des performances de séparation statistiquement significative ;
- enregistrée dans des conditions acoustiques différentes allant de la chambre anéchoïque à la pièce avec une réverbération significative et ceci afin de tester la robustesse des algorithmes de séparation face à la réverbération ;
- variée afin de tester la capacité des algorithmes de séparation à généraliser. Pour une base de données de parole, elle doit contenir des extraits de parole enregistrés avec des voies différentes d'hommes et de femmes, les sources audio doivent provenir de directions d'arrivées différentes et variées.

Enregistrer directement avec le réseau de capteurs des voix d'hommes et de femmes dans des conditions acoustiques différentes et des situations variées prend beaucoup de temps et nécessite un grand nombre de locuteurs volontaires et un nombre plus grand de mesures.

Pour construire nos bases de données de parole dans des conditions réelles sans avoir à répéter les enregistrement avec les locuteurs volontaires, nous considérons d'abord une base de données de parole enregistrée dans des conditions anéchoïques, c'est à dire sans réverbération ni bruit, nous appelons cette base de données Sources-pures. Les mélanges convolutifs sont construits comme suit :

1. dans chaque condition acoustique (pièce), estimer les réponses impulsionnelles acoustique $h_{ij}(l)$ de différents points d'émission $1 \leq i \leq N$ (différentes directions d'arrivées) vers chacun des microphones du réseau de capteurs $1 \leq j \leq M$;
2. calculer les contributions aux capteurs $[s_{ij}(t)]_{1\leq i\leq N,\, 1\leq j\leq M}$ pour chacune des sources
   $[s_i(t)]_{1\leq i\leq N}$ considérées et des points d'émission voulu ; pour chaque capteur j

et chaque source i, la contribution $s_{ij}(t)$ est la convolution d'une source pure $s_i(t)$ avec la réponse impulsionnelle acoustique d'un point d'émission choisi et le capteur j, ceci donne : $s_{ij}(t) = \sum_{l=0}^{L-1} h_{ij}(l)\, s_i(t-l)$;

3. pour chaque capteur, le signal reçu consiste en la somme des N sources images calculées au niveaux de ce capteur : $x_j(t) = \sum_{i=0}^{N} s_{ij}(t)$.

Pour chaque pièce, les réponses impulsionnelles acoustiques sont calculées une seule fois pour chacun des points sources considérés vers chaque capteur, nous pouvons ensuite changer de locuteur hors enregistrement et autant de fois que l'on veut en utilisant la base de données des sources pures.

7.4 Calcul des réponses impulsionnelles avec les séquences complémentaires de Golay

Nous considérons un système à temps discret caractérisé par sa réponse impulsionnelle $h(t)$, un signal d'entrée $s(t)$ et un signal de sortie $x(t)$, $t \in \mathbb{Z}$. Pour identifier le système, nous avons besoin d'estimer $h(t)$ pour un signal d'entrée $s(t)$ connu et un signal de sortie $x(t)$. Dans notre cas, le système représente le chemin acoustique entre un point d'émission et un capteur et nous voulons estimer la réponse impulsionnelle acoustique $h(t)$:

$$x(t) = s(t) * h(t) \tag{7.2}$$

où $*$ est un opérateur de convolution.

Pour estimer cette réponse impulsionnelle, nous utilisons les séquences complémentaires de Golay [33] comme signal d'entrée. Les séquences complémentaires de Golay ont la propriété intéressante que leur fonctions d'autocorrélation ont des lobes secondaires complémentaires : la somme des fonctions d'autocorrélation est nulle partout sauf à l'origine. Les séquences complémentaires de Golay ne sont pas uniques [34]. Nous utilisons une paire de séquences $a(t)$ et $b(t)$ de longueur L et définie comme suit :

$$\begin{aligned} a(t) &= \pm 1 \text{ pour } 1 \leq t \leq L \\ b(t) &= \pm 1 \text{ pour } 1 \leq t \leq L \end{aligned} \tag{7.3}$$

Ces séquences sont des séquences complémentaires de Golay si et seulement si :

$$a(-t) * a(t) + b(-t) * b(t) = 2L\delta(t) \tag{7.4}$$

où $\delta(t)$ est l'impulsion de Dirac. Les séquences que nous utilisons sont définies récursivement comme suit :

$$\begin{bmatrix} A_L \\ B_L \end{bmatrix} = \begin{bmatrix} A_{L/2} & B_{L/2} \\ A_{L/2} & -B_{L/2} \end{bmatrix} \text{ avec } \begin{matrix} A_L = [a(1) \cdots a(L)] \\ B_L = [b(1) \cdots b(L)] \end{matrix}$$

$$\text{et } \begin{bmatrix} A_2 \\ B_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

Les réponses du système aux entrées $a(t)$ et $b(t)$ sont respectivement :

$$\begin{aligned} x_a(t) &= a(t) * h(t) \\ x_b(t) &= b(t) * h(t) \end{aligned} \tag{7.5}$$

En utilisant les équations (7.4) et (7.5), la réponse impulsionnelle du système est donnée par :

$$h(t) = \frac{1}{2L}\left(a(-t) * x_a(t) + b(-t) * x_b(t)\right) \tag{7.6}$$

Dans un cas pratique, nous utilisons comme entrée une séquence de Golay construite comme le montre la figure 7.5. Chaque séquence de Golay A et B de longueur L est répétée N_G fois. Après la mesure, le premier bloc de chacune des deux séries qui sont répétées n'est pas pris en compte dans l'estimation de la réponse impulsionnelle, car ce premier bloc contient une convolution avec le bloc précédent, qui pour la première série était un bloc nul, et pour la seconde série était un bloc de la première série. Pour le reste des blocs mesurés $x_A = [x_a(1) \ldots x_a(L)]$ et $x_B = [x_b(1) \ldots x_b(L)]$, nous estimons les réponses impulsionnelles dans le domaine fréquentiel. La réponse impulsionnelle finale est la moyenne des réponses impulsionnelles estimées. Cette répétition assure une certaine robustesse dans la mesure de la réponse.

Cette méthode de calcul des réponses impulsionnelles a été utilisée pour le calcul des réponses impulsionnelles acoustiques dans différentes pièces réverbérantes et pour l'estimation des réponses impulsionnelles de tête (HRIR : Head Related Transfer Functions) utilisées pour calculer les fonctions de transfert de tête comme nous l'avons présenté dans la section 3.4. Dans la suite, nous présenterons le détail de ces mesures.

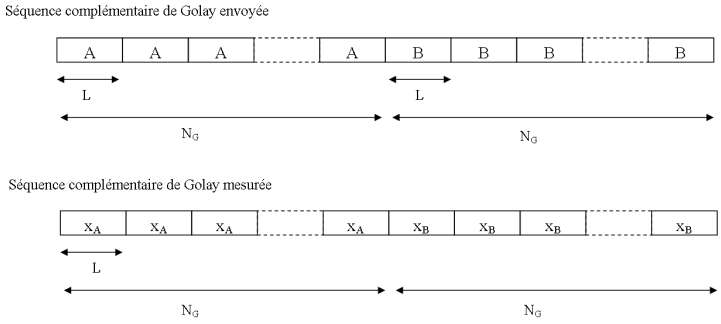


Figure 7.5 – Structure totale des séquences complémentaires de Golay à l'entrée et celles mesurées

7.5 Base de données des réponses impulsionnelles dans un milieu réverbérant

7.5.1 A Télécom ParisTech

Nous avons enregistré une première base de réponses impulsionnelles dans le studio de Télécom ParisTech dont le taux de réverbération est de RT_{60}= 300 ms, calculé à partir de la courbe de décroissance d'énergie 7.4b. Nous avons estimé les réponses impulsionnelles de différentes positions comme le montre la figure 7.6. Les points d'émission sont placés à 1m20 du réseau de capteurs et nous avons mesuré les réponses impulsionnelles de -90° à 90° avec un pas de 10° (à l'exception de -10° et 10° où nous n'avons pas fait de mesure).

Nous avons choisi d'évaluer les algorithmes sur un cas de séparation de 2 et 3 sources. Pour un cas de séparation de deux sources, la première source est placée à 0° et la seconde source est choisie entre 20° et 90°. Pour un cas de séparation de trois sources, la première source est toujours placée à 0° et la $2^{ème}$ et $3^{ème}$ source sont choisies entre -90° et -20° et entre 20° et 90°. Nous appelons ces bases de données Theo-RI-studio.

7.5.2 A l'institut de la vision

Dans le cadre du projet Romeo, un appartement témoin a été conçu à l'institut de la vision. Cet appartement est une vitrine pour le projet Romeo, c'est un appartement témoin qui sert à l'acquisition de base de données commune aux différents partenaires du projet et à faire les premiers tests de Romeo dans un environnement proche de celui dans lequel l'humanoïde évoluera. Dans cet appartement témoin,

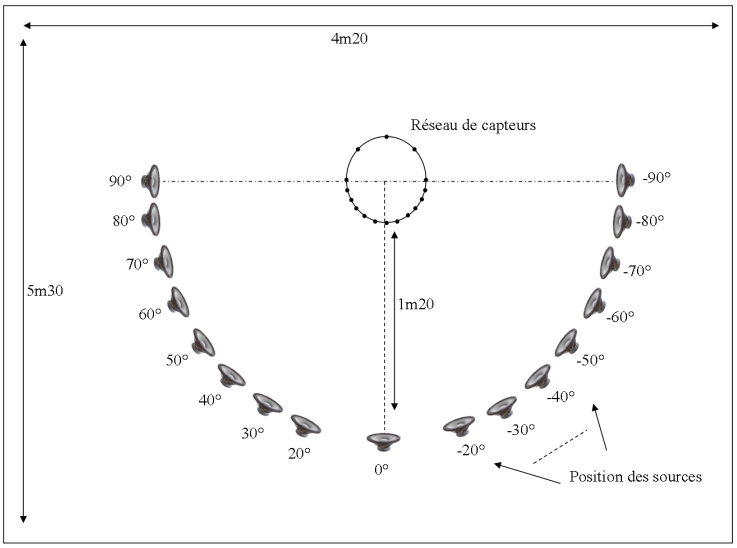


FIGURE 7.6 – Position des sources par rapport au réseau de capteurs dans le studio de Télécom ParisTech

nous avons procédé à des mesures de réponses impulsionnelles de trois angles d'arrivée différents : 20°, 60° et -35° comme montré à la figure 7.7. Comme l'indique la courbe de décroissance de l'énergie 7.4c, le temps de réverbération de cet appartement témoin de l'institut de la vision est de RT_{60}= 600 ms. Nous appelons ces bases de données Theo-RI-IDV.

7.6 Base de données de HRTF

Nous rappelons qu'une fonction de transfert de tête ou HRTF est la représentation fréquentielle d'une réponse impulsionnelle de tête ou HRIR qui caractérise comment un signal émis d'une direction spécifique est reçu à une oreille. La HRIR de chaque oreille capture l'information de localisation du signal source et l'altération produite par la tête et le pavillon sur le champ sonore proche [16]. Nous étendons le principe des HRIR et HRTF à la problématique de l'audition des robots avec un réseau de capteurs, donc avec plus de deux "oreilles", nous mesurons les HRIR et HRTF pour les 16 capteurs de Theo.

A notre connaissance, ceci représente un premier cas d'une base de données de HRTF et HRIR multicapteurs avec différents angles d'azimut et d'élévation ; jusqu'à maintenant, les bases de données disponibles pour le téléchargement sont binaurales [12, 27, 28, 35]. KEMAR est une base de données binaurale mesurée avec une tête

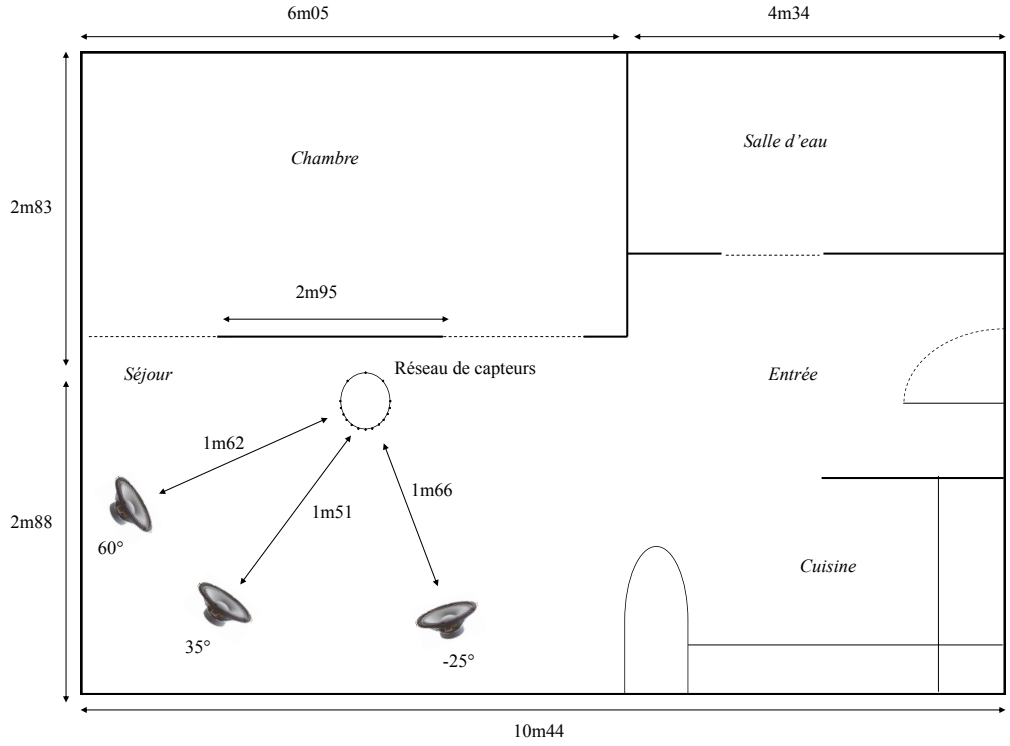


Figure 7.7 – Position des sources par rapport au réseau de capteurs dans l'Institut De la Vision (IDV)

de mannequin pour 70 positions [35] et la base de données de HRTF CIPIC a été mesurée avec 45 sujets humains pour 25 azimuts et 50 élévations [12].

Le calcul des HRIR se fait en utilisant les séquences complémentaires de Golay, dans la suite la description du processus de mesure.

7.6.1 Description matérielle et logicielle

L'estimation des réponses impulsionnelles de tête en utilisant les séquences complémentaires de Golay a été faite dans la salle anéchoïque de Télécom ParisTech comme le montre la figure 7.8. Nous avons utilisé le matériel suivant :

- Theo pour l'acquisition de la base de données des HRIR et HRTF : Theo-HRTF ;
- deux enregistreurs Echo Audiofire Pre8 [1] ;

1. http://www.echoaudio.com/Products/FireWire/AudioFirePre8/index.php

– 16 microphones AKG C417 pp [2] (*cf.* figure 11.2) ;
– 7 haut-parleurs Tannoy System 600 [3] ;
– une table tournante Brüel & Kjær Type 9640 [4].

Les enregistreurs Audiofire Pre8 sont reliés par FireWire à un PC ayant comme système d'exploitation Linux avec un noyau temps réel. Les logiciels utilisés sont :

– Le kit de connexion audio Jack (JACK-control) [46] une application open-source qui contrôle le serveur son spécifique à l'infrastructure Linux Audio Desktop.
– Ffado [32], un pilote open-source pour les dispositifs pro-audio se basant sur des connexions par FireWire. Nous utilisons ffado-mixer pour contrôler la synchronisation entre les enregistreurs et accéder à la table de mixage des canaux des enregistreurs.

7.6.2 Processus expérimental

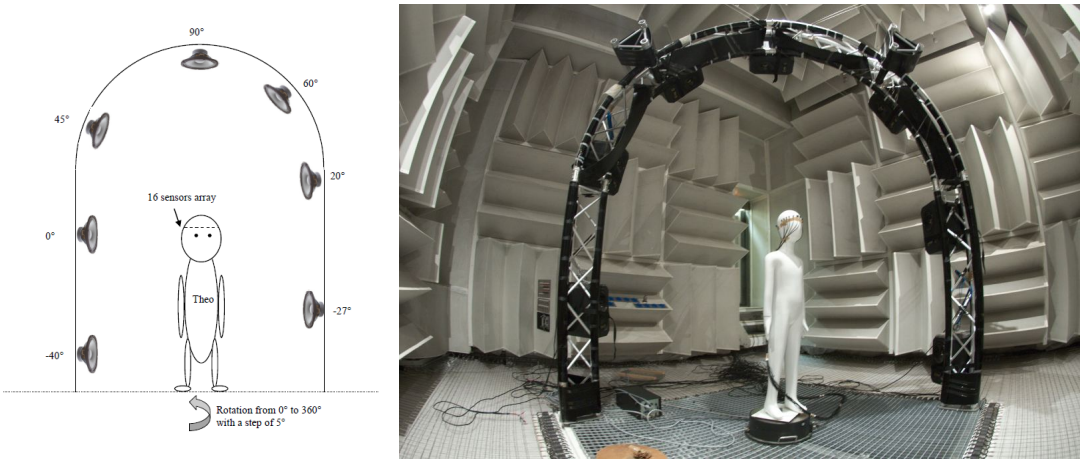


FIGURE 7.8 – Theo et la position des hauts-parleurs pour l'enregistrement des séquences complémentaires de Golay dans la chambre anéchoïque de Télécom ParisTech

Dans la chambre anéchoïque, nous avons mesuré des HRTF depuis 504 points d'azimut et d'élévation distribués comme suit (*cf.* figure 7.8) :

– 73 angles d'azimut de 0° à 355° avec un pas de 5°
– 7 angles d'élévation : -40°, -27°, 0°, 20°, 45°, 60° et 90°

2. http://www.akg.com/mediendatenbank2/psfile/datei/35/c4174055c447d8838.pdf
3. http://www.tannoy.com/products/158/uman_System600.pdf
4. http://www.bksv.com/products/Télécomaudiosolutions/electroacousticsaccessories/turntablesystemtype9640.aspx

Theo est fixé sur une table tournante dans le centre de l'arc supportant les haut-parleurs. Une séquence complémentaire de Golay est émise séquentiellement de chaque haut-parleur (chaque élévation) et enregistrée avec un réseau de capteurs de 16 microphones pour chaque angle d'azimut. Cette base de données Theo-HRTF est disponible avec les fréquences d'échantillonnage de 16kHz et 48kHz et peut être téléchargée par ce lien `http://www.tsi.telecom-paristech.fr/aao/?p=347`

Conclusion

Nous avons présenté dans le chapitre les différentes bases de données que nous avons acquises avec une configuration de réseaux de 16 capteurs placés autour de la tête d'un mannequin de vitrine Theo. La base de données des HRTF sert à calculer les filtres de formation de voies utilisés dans l'étape de prétraitement de nos algorithmes de séparation à deux étapes. La base de données de mélanges convolutifs, acquise dans deux milieux différents, sert à évaluer et comparer les algorithmes de séparation de sources proposés. Le tableau 7.1 présente un récapitulatif de ces bases de données.

Base de données	Nombre de données
Sources-pures	40 paires
Theo-RI-studio	320 paires (de 0° à -90°)
Theo-RI-IDV	120 paires
Theo-HRTF	504 HRIR/HRTF

Table 7.1 – Récapitulatif des bases de données enregistrées avec Theo (16 capteurs chacun)

Chapitre 8

Outils d'évaluation des algorithmes de séparation de sources

Introduction

L'évaluation ou la comparaison des performances des algorithmes de séparation de sources a pour longtemps été une tâche difficile. En effet, il y avait un manque de méthodes de mesure des performances et celles qui existaient présentaient de nombreuses limitations [88]. Prenons l'exemple de la différence inter-symbole entre la source estimée et la vraie source ou la comparaison directe de la source estimée et la source originale : ces deux méthodes ne considèrent que le problème de permutation et d'échelle et ne prennent pas en compte les distorsions autorisées par l'application, de plus ces méthodes ne présentent qu'un seul type de critère de performance englobant toutes les erreurs.

En 2006, Vincent *et al.* [88] proposent une méthode *objective* d'évaluation des performances des algorithmes de séparation de sources basée sur quatre types de distorsions autorisées, allant de la distorsion la plus simple du gain invariant temporellement à la plus complexe des filtres variables dans le temps. Dans chaque cas, ils décomposent la source estimée en la vraie source plus des termes d'erreurs correspondant aux interférences, au bruit additif et aux artéfacts introduits par l'algorithme de séparation. Ensuite, ils dérivent des mesures de performance globale en utilisant le rapport d'énergie et une mesure de performance séparée pour chaque terme d'erreur. Ces mesures de performance sont implémentées dans une boite à outils MATLAB sous le nom de BSS_EVAL distribuée sous la licence publique générale GNU.

Cependant, les algorithmes de décomposition de la distorsion proposés jusqu'à maintenant ne donnent pas toujours les composantes attendues et on peut mettre

en doute la capacité des rapports des énergies à correspondre aux scores *subjectifs* puisque des phénomènes audio tels que la perception de la "loudness" et le masquage spectral ne sont pas pris en compte. Emiya *et al.* [30] proposent d'évaluer *la qualité perçue* des signaux sources estimées dans le contexte de la séparation de sources audio. Ces signaux peuvent impliquer un ou plusieurs types de distorsions, y compris les distorsions de la source cible, les interférences provenant des autres sources ou les artéfacts du bruit musical. Ils proposent un protocole de test subjectif pour évaluer la qualité perçue par rapport à chaque type de distorsion, et une famille de mesures objectives afin de prédire ces scores subjectifs. Cette série de mesures objectives est basée sur la décomposition de l'erreur estimée en différentes composantes de distorsions et l'utilisation de l'outil de mesure perceptuelle PEMO-Q. Cette méthode d'évaluation réalisée avec MATLAB est disponible en téléchargement libre sous le nom de PEASS (Perceptual Evaluation methods for Audio Source Separation).

Pour l'évaluation de nos algorithmes de séparation de sources, nous utilisons les deux méthodes BSS_EVAL et PEASS que nous détaillerons dans la suite, elles ont été utilisées aussi dans la campagne d'évaluation de séparation des signaux SISEC (Signal Separation Evaluation Campaign). Pour ces méthodes d'évaluation, la connaissance des sources originales est nécessaire. De plus, ces mesures ne tiennent pas compte du problème de permutation de la séparation aveugle de sources audio : si nécessaire, la source estimée peut être comparée à toutes les sources originales et la "vraie source" peut être sélectionnée comme celle donnant les meilleurs résultats.

8.1 Évaluation objective des performances de séparation (BSS_EVAL)

Les mesures de performance sont calculées pour chaque source estimée en la comparant avec la source originale. Le calcul du critère se fait en deux étapes : la décomposition de la source estimée et le calcul des différents rapports d'énergies qui donnent les mesures de performances globales. On note le vecteur temporel relatif à une source j et de longueur T : $\mathbf{x}_j = [x_j(1), \ldots, x_j(T)]^T$.

8.1.1 Décomposition de la source estimée

La première étape consiste à décomposer la source estimée $\mathbf{y}_j$ comme suit :

$$\mathbf{y}_j = \mathbf{s}_j^{\text{target}} + \mathbf{e}_j^{\text{interf}} + \mathbf{e}_j^{\text{noise}} + \mathbf{e}_j^{\text{artif}} \tag{8.1}$$

où $\mathbf{s}_j^{\text{target}} = f(\mathbf{s}_j)$ est une version de la vraie source $\mathbf{s}_j$ modifiée par une distorsion autorisée $f \in \mathcal{F}$, dans notre cas la distorsion consiste en un filtrage temporel invariant, et $\mathbf{e}_j^{\text{interf}}$, $\mathbf{e}_j^{\text{noise}}$ et $\mathbf{e}_j^{\text{artif}}$ sont respectivement les termes d'erreur relatifs aux *interférences*, au *bruit* et aux *artéfacts*. Ces quatre termes doivent représenter la partie de $\mathbf{y}_j$ perçue comme venant de la source désirée $\mathbf{s}_j$, des sources non désirées $(\mathbf{s}_{j'})_{j' \neq j}$, du bruit des capteurs $(\mathbf{n}_i)_{1 \leq i \leq M}$ et d'autres causes comme les distorsions non autorisés et/ou le bruit musical). Quand une distorsion par filtrage temporel invariant est autorisée, $\mathbf{s}_j^{\text{target}}$ est une version filtrée de $\mathbf{s}_j$ telle que $s_j^{\text{target}}(t) = \sum_{l=0}^{L-1} h(l) s_j(t-l)$. Par conséquent, $\mathbf{s}_j^{\text{target}}$ appartient au sous-espace engendré par des versions décalées de $\mathbf{s}_j$, ce qui implique que $\mathbf{s}_j^{\text{target}}$ peut être définie en projetant $\mathbf{y}_j$ sur ce sous-espace.

Soit $\Pi\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ le projecteur orthogonal sur le sous-espace engendré par les vecteurs $\mathbf{x}_1, \ldots, \mathbf{x}_k$. Ce projecteur est une matrice $T \times T$ où T est la longueur de ces vecteurs. On note $\mathbf{s}_j^l$ et $\mathbf{n}_i^l$ le signal source $\mathbf{s}_j$ et le signal bruit $\mathbf{n}_i$ décalés par l échantillons, ce qui donne : $s_j^l(t) = s_j(t-l)$ et $n_i^l(t) = n_i(t-l)$. Soient les trois projecteurs suivant :

$$P_{\mathbf{s}_j} = \Pi\left\{\left(\mathbf{s}_j^l\right)_{0 \leq l \leq L-1}\right\} \tag{8.2}$$

$$P_{\mathbf{s}} = \Pi\left\{\left(\mathbf{s}_{j'}^l\right)_{0 \leq j' \leq N, 0 \leq l \leq L-1}\right\} \tag{8.3}$$

$$P_{\mathbf{s},\mathbf{n}} = \Pi\left\{\left\{\left(\mathbf{s}_{j'}^l\right)_{0 \leq j' \leq N}, \left(\mathbf{n}_i^l\right)_{0 \leq i' \leq M}\right\}_{0 \leq l \leq L-1}\right\} \tag{8.4}$$

La source estimée y_j se décompose alors comme la somme des quatre termes :

$$\mathbf{s}_j^{\text{target}} = P_{\mathbf{s}_j}\mathbf{y}_j \tag{8.5}$$

$$\mathbf{e}_j^{\text{interf}} = P_{\mathbf{s}}\mathbf{y}_j - P_{\mathbf{s}_j}\mathbf{y}_j \tag{8.6}$$

$$\mathbf{e}_j^{\text{noise}} = P_{\mathbf{s},\mathbf{n}}\mathbf{y}_j - P_{\mathbf{s}}\mathbf{y}_j \tag{8.7}$$

$$\mathbf{e}_j^{\text{artif}} = \mathbf{y}_j - P_{\boldsymbol{s},\mathbf{n}}\mathbf{y}_j \tag{8.8}$$

8.1.2 Mesures des performances globales

Dans la deuxième étape, les rapports des énergies en décibel sont calculés afin d'évaluer le taux de participation de chacun de ces quatre termes à la source estimée $\mathbf{y}_j$, soit dans toute la durée du signal soit sur des trames locales :

– le rapport source-à-distorsion SDR (Signal-to-Distorsion Ratio) :

$$\mathrm{SDR} = 10\log_{10}\frac{\left\|\mathbf{s}_j^{\mathrm{target}}\right\|^2}{\left\|\mathbf{e}_j^{\mathrm{interf}}+\mathbf{e}_j^{\mathrm{noise}}+\mathbf{e}_j^{\mathrm{artif}}\right\|^2} \tag{8.9}$$

– le rapport source-à-interférences SIR (Sources-to-Interferences Ratio) :

$$\mathrm{SIR} = 10\log_{10}\frac{\left\|\mathbf{s}_j^{\mathrm{target}}\right\|^2}{\left\|\mathbf{e}_j^{\mathrm{interf}}\right\|^2} \tag{8.10}$$

– le rapport sources-sur-bruit SNR (Sources-to-Noise Ratio) :

$$\mathrm{SNR} = 10\log_{10}\frac{\left\|\mathbf{s}_j^{\mathrm{target}}+\mathbf{e}_j^{\mathrm{interf}}\right\|^2}{\left\|\mathbf{e}_j^{\mathrm{noise}}\right\|^2} \tag{8.11}$$

– le rapport sources-à-artéfacts SAR (Sources-to-Artifacts Ratio) :

$$\mathrm{SAR} = 10\log_{10}\frac{\left\|\mathbf{s}_j^{\mathrm{target}}+\mathbf{e}_j^{\mathrm{interf}}+\mathbf{e}_j^{\mathrm{noise}}\right\|^2}{\left\|\mathbf{e}_j^{\mathrm{artif}}\right\|^2} \tag{8.12}$$

Pour l'évaluation de nos algorithmes, nous évaluons le SDR, le SIR et le SAR. Le SNR ne sera pas pris en compte car nous considérons un bruit diffus spatialement décorrélé dont l'énergie est supposée être negligeable par rapport à celle des sources. Si le bruit est ponctuel, il sera considéré comme une source sonore.

8.2 Évaluation perceptuelle des performances de séparation (PEASS)

Emiya *et al.* [30] proposent une famille de mesures objectives basée sur une décomposition de la distorsion différente de celle de BSS_EVAL et valident la capacité de ces mesures objectives à prédire les scores subjectifs multi-critères obtenus lors d'un protocole de test dédié à évaluer la qualité de la séparation de sources audio. Les critères subjectifs considérés dans l'évaluation perceptuelle de la qualité de la

séparation de sources audio sont :

1. score global : la *qualité globale* par rapport à la référence pour chaque signal test (chaque source estimée) ;
2. préservation de la source cible : la qualité en termes de *préservation de la source cible* dans chaque signal test ;
3. suppression des autres sources : la qualité en termes de *suppression des autres sources* dans chaque signal test ;
4. addition de bruit artificiel : la qualité en termes d'*absence de bruit artificiel additionnel* dans chaque signal test.

Dans ce qui suit, nous présenterons la famille des quatre mesures objectives dont le but est de prédire les scores subjectifs que nous venons de citer. L'approche proposée par [30] consiste à diviser le signal de distorsion en une somme de composantes liées à la *distorsion cible*, aux *interférences* et aux *artéfacts*, à évaluer leur saillance perceptuelle en utilisant des métriques de nature auditives et à combiner les attributs résultants.

8.2.1 Modélisation et estimation des composantes de distorsions

Le distorsion entre la source estimée $y_j(t)$ et la source cible $s_j(t)$ est décomposée en la somme d'une composante de distorsion cible $e_j^{\text{target}}(t)$, une composante d'interférence $e_j^{\text{interf}}(t)$ et une composante d'artéfact $e_j^{\text{artif}}(t)$ comme suit :

$$y_j(t) - s_j(t) = e_j^{\text{target}}(t) + e_j^{\text{interf}}(t) + e_j^{\text{artif}}(t) \tag{8.13}$$

Pour accomplir cette décomposition, on doit spécifier comment la distorsion cible et les composantes d'interférences sont liées aux sources originales. Cependant, la manière dont le système auditif sépare les flux associés à ces composantes demeure inconnue. Une approche consiste à supposer que ces composantes sont des versions linéairement distordues des sources réelles et cette distorsion est modélisée par un filtre à réponse impulsionnelle finie (FIR) multicanal invariant dans le temps. Cette hypothèse a été prise en compte notamment dans BSS_EVAL. Cependant, ces composantes de distorsion ne correspondent pas toujours à celles perçues par l'oreille humaine. Ceci est dû en particulier au modèle invariant dans le temps qui ne correspond pas à la nature variable dans le temps des distorsions rencontrées et à la résolution fréquentielle constante des filtres RIF qui ne correspond pas à celle de

l'oreille. Une décomposition à temps-variable a été proposée par [88]. Cependant, à cause de son grand coût de calcul, elle est restreinte en pratique aux filtres avec une basse résolution spatiale et temporelle, et par conséquent, elle n'a pas amélioré les résultats. La décomposition proposée par Emiya *et al.* [30] a pour but de résoudre ces problèmes et donne des composantes de distorsion perceptuellement plus pertinentes s'approchant de la résolution temps-fréquence auditive, grâce à l'utilisation de banc de filtre. Ceci se fait en trois étapes :

1. analyse temps-fréquence : la source estimée $y_j(t)$ et les sources originales $s_i(t)$, $1 \leq i, j \leq N$ sont partitionnés en temps et en fréquence par un banc de filtres type gammatone [1] en des signaux $y_{ib}(t)$ et $s_{ib}(t)$ indexés par b. Dans chaque sous bande, après une étape de sous-échantillonnage, ces signaux sont ensuite fenêtrés en des trames recouvrantes indexées par u : $y_{jbu} = w_a(t)y_{jb}(t - uN)$ et $s^{\tau}_{ibu} = w_a(t)s_{ib}(t - uN - \tau)$, où w_a est la fenêtre d'analyse, N est le pas d'avancement et $s_{ib}(t-\tau)$ est la version décalée de la vraie source $s_{ib}(t)$, $-L/2 \leq \tau \leq L/2$;

2. décomposition par moindres carrés jointe : à cause de la large bande passante des filtres gammatone, les composantes de distorsion sont estimées par un filtrage additionnel en chaque sous-bande et trame temporelle ; ces composantes sont définies par un filtrage RIF multicanal invariant dans le temps des sources cibles et des sources interférentes ; les coefficients de ces filtres sont estimés par une projection des moindres carrés de la distorsion $y_{jbu}(t) - s^0_{jbu}(t)$ sur le sous-espace engendré par les versions décalées des signaux sources $s^{\tau}_{ibu}(t)$, $1 \leq i \leq N$ et $-L/2 \leq \tau \leq L/2$; les composantes de distorsion s'écrivent :

$$e^{\text{target}}_{jbu}(t) = \sum_{\tau=-L/2}^{L/2} \alpha_{jbu,j}(\tau)\, s^{\tau}_{jbu}(t) \tag{8.14}$$

$$e^{\text{interf}}_{jbu}(t) = \sum_{i \neq j} \sum_{\tau=-L/2}^{L/2} \alpha_{jbu,l}(\tau)\, s^{\tau}_{ibu}(t) \tag{8.15}$$

$$e^{\text{artif}}_{jbu}(t) = y_{jbu}(t) - s^0_{jbu}(t) - e^{\text{target}}_j(t) - e^{\text{interf}}_j \tag{8.16}$$

3. reconstruction des signaux temporels : les signaux sont reconstruit par overlap-add et inversion du banc de filtre $e^{\text{target}}_j(t)$, $e^{\text{interf}}_j(t)$ et $e^{\text{artif}}_j(t)$.

1. Un banc de filtres gammatone est un banc de filtres qui modélise la non-linéarité et la variance temporelle du système auditif.

8.2.2 Mesures objectives

Étant données des composantes de distorsion comme celles calculées à la sous-section précédente ou comme celles utilisées dans BSS_EVAL, le but est d'évaluer la similarité entre la source estimée et la source originale en s'appuyant sur les quatre critères subjectifs cités dans l'introduction de cette section. Ceci est fait en deux étapes.

La première étape consiste à évaluer perceptuellement chaque composante de distorsion en utilisant la métrique de similarité perceptuelle PSM (Perceptual Similarity Measure) fournit par le modèle auditif PEMO-Q [47].

La métrique de similarité perceptuelle donne les attributs suivant :

$$q_j^{\text{overall}} = \text{PSM}\left(\mathbf{y}_j, \mathbf{s}_j\right) \tag{8.17}$$

$$q_j^{\text{target}} = \text{PSM}\left(\mathbf{y}_j, \mathbf{y}_j - e_j^{\text{target}}\right) \tag{8.18}$$

$$q_j^{\text{interf}} = \text{PSM}\left(\mathbf{y}_j, \mathbf{y}_j - e_j^{\text{interf}}\right) \tag{8.19}$$

$$q_j^{\text{artif}} = \text{PSM}\left(\mathbf{y}_j, \mathbf{y}_j - e_j^{\text{artif}}\right) \tag{8.20}$$

Dans la deuxième étape, une combinaison non linéaire de ces métriques donne les quatre mesures objectives suivantes [30] :

- Le score perceptuel global OPS (Overall Perceptual Score) qui évalue la qualité globale de la source estimée par rapport à la source originale.
- Le score perceptuel relatif à la cible TPS (Target-related Perceptual Score) qui traduit la qualité en termes de préservation de la source cible dans la source estimée.
- Le score perceptuel relatif aux interférences IPS (Interference-related Perceptual Score) qui évalue la qualité en termes de suppression des autres sources dans la source estimée.
- Le score perceptuel relatif aux artéfacts APS (Artifacts-related Perceptual Score) qui estime la qualité en termes d'absence de bruit artificiel additionnel dans chaque source estimée.

Conclusion

Dans ce chapitre, nous avons introduit les outils d'évaluation des performances des algorithmes de séparation de sources que nous avons utilisés pour évaluer nos algorithmes. Dans un premier temps, nous avons présenté la boite à outils BSS_EVAL

[88] contenant des mesures objectives de la qualité des sources estimées : le rapport source-à-interférence (SIR), le rapport source-à-distorsion (SDR), le rapport sources-à-artéfacts (SAR) et le rapport sources-sur-bruit (SNR). Parfois, ces mesures objectives sont insuffisantes pour évaluer la qualité perçue des sources estimées. Par conséquent, nous avons introduit dans un deuxième temps la boite à outils PEASS [30] qui propose une évaluation perceptuelle des performances des algorithmes de séparation de sources selon les scores suivants : le score perceptuel global (OPS), le score perceptuel relatif à la cible (TPS), le score perceptuel relatif aux interférences (IPS) et le score perceptuel relatif aux artéfacts (APS).

Chapitre 9

Évaluation des algorithmes itératifs de séparation de sources

Introduction

Dans ce chapitre, nous présentons les performances des algorithmes itératifs de séparation de sources présentés dans les chapitres 4 et 5 à savoir :

- l'algorithme de séparation basé sur un critère de parcimonie BSS-l_1 que nous comparons à l'analyse en composantes indépendantes ICA ;
- l'algorithme de séparation basé sur un critère de parcimonie variable BSS-l_p qui se traduit par la minimisation de la norme l_p paramétrée ;
- la classe d'algorithmes avec prétraitement de formation de voies avec les quatre différentes configurations de ce prétraitement à savoir :
  - formation de N voies vers les directions d'arrivées des sources BF_DOA ;
  - formation de K voies fixe BF_fixed ;
  - formation de voies fixe avec sélection des lobes ayant la plus grande énergie BF_fixed_BS ;
  - formation de voies fixe avec estimation du nombre de source BF_fixed_NbSrEstim.

Dans la suite, nous détaillerons les paramètres de ces algorithmes de séparation itératifs et nous procèderons à une évaluation de leurs performances en utilisant les outils d'évaluation présentés dans la section précédente : BSS_EVAL et PEASS.

Paramètres des algorithmes de séparation itératifs

Les signaux de test des différentes bases de données ont une durée de 5s et sont échantillonnés à 16 kHz. La fenêtre d'analyse spectrale est de type Hanning et de longueur 128 ms (2048 échantillons) et le pas d'avancement est de 50%. Le pas de mise à jour des algorithmes itératifs est $\mu = 0.2$ choisi d'après l'état de l'art [29]. Nous choisissons un nombre d'itérations égal à 100 et nous verrons que les algorithmes présentés convergent bien avant d'atteindre ce nombre.

9.1 Comparaison entre le critère de parcimonie et le critère d'indépendance

Nous commençons cette évaluation des algorithmes de séparation de sources proposés par la comparaison des performances de BSS-l_1, l'algorithme basé sur un critère de séparation de parcimonie qui est la minimisation de la norme l_1 ; et de l'algorithme basé sur un critère d'indépendance qui est l'analyse en composantes indépendantes avec minimisation de l'information mutuelle de Douglas et Gupta [29].

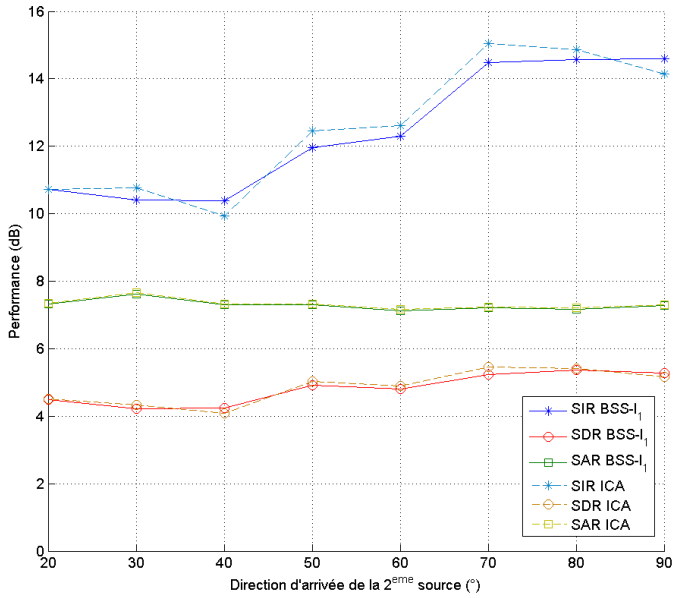


FIGURE 9.1 – Le rapport source-à-interférences SIR (courbes en bleu), le rapport source-à-distorsion SDR (courbes en rouge) et le rapport sources-à-artéfacts SAR (courbes en vert) des algorithmes BSS-l_1 (courbes continues) et ICA (courbes en tirets) : évaluation sur la base de données Theo-RI-studio.

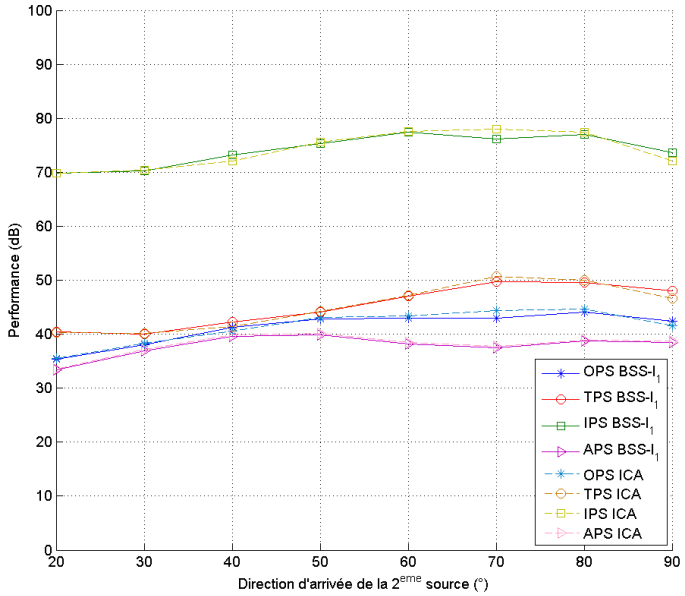
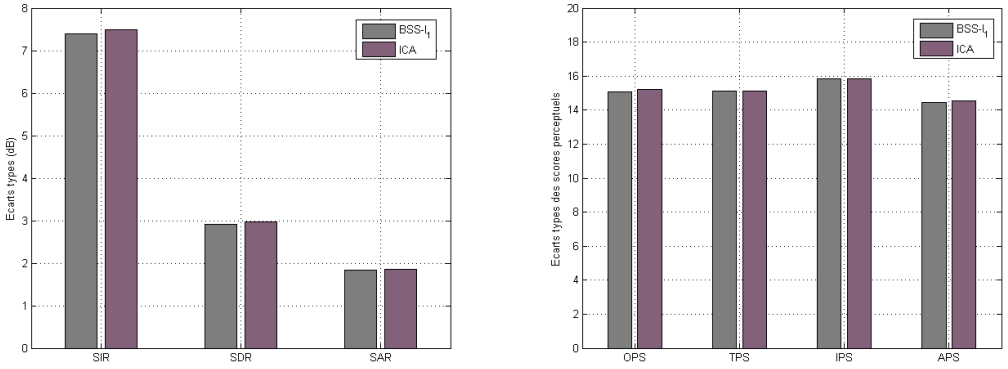


FIGURE 9.2 – Le score perceptuel global OPS (courbes en bleu), le score perceptuel relatif à la cible TPS (courbes en rouge), le score perceptuel relatif aux interférences IPS (courbes en vert) et le score perceptuel relatifs aux artéfacts APS (courbes en mauve) des algorithmes BSS-l_1 (courbes continues) et ICA (courbes en tirets) : évaluation sur la base de données Theo-RI-studio.



(a) Les écarts types des SIR, SDR et SAR (b) Les écarts types des OPS, TPS, IPS et ASP

FIGURE 9.3 – Les écarts types des résultats de séparations des algorithmes BSS-l_1 (barres grises) et ICA (barres violettes) : évaluation sur la base de données Theo-RI-studio.

La figure 9.1 montre une comparaison du rapport signal-sur-bruit SIR, le rapport signal-à-distorsion SDR et le rapport signal-à-artéfacts SAR de ces deux algorithmes évalués sur la base de données Theo-RI-studio. Les courbes montrent que les performances de ces deux approches sont comparables. Nous retrouvons ce même constat dans les performances de l'évaluation perceptuelle du résultat de séparation de ces deux algorithmes (*cf.* figure 9.2). En effet, le score perceptuel global OPS, le score perceptuel relatif à la cible TPS, le score perceptuel relatif aux interférences IPS et le score perceptuel relatif aux artéfacts APS de BSS-l_1 et d'ICA sont proches comme le confirme la figure 9.3 des écarts types. Ceci peut s'expliquer par le fait que l'analyse en composantes indépendantes conduit à la minimisation de la parcimonie des sources à estimer à partir des mélanges convolutifs, et ceci de la même façon que la minimisation de la norme l_1.

9.2 Minimisation de la norme l_p paramétrée

Avant le développement de l'algorithme basé sur la minimisation de la norme l_p paramétrée BSS-l_p-param [5], nous avons essayé de savoir comment se comporte un algorithme basé sur la minimisation de la norme l_p avec différentes valeurs du paramètre p, donc différentes contraintes de parcimonie. Nous avons considéré un paramètre p entre 0.1 (contrainte de parcimonie la plus dure) et 1 (contrainte de parcimonie la moins dure). La figure 9.4 montre le résultat de la séparation en termes de rapport signal-à-interférences de quatre cas de séparation de deux sources. Nous remarquons que pour la pseudo norme l_p avec un paramètre $p < 1$, nous pouvons avoir de meilleurs résultats de séparation que pour la norme l_1. Cependant, la valeur du paramètre p pour laquelle le SIR est le plus élevé est différente d'un cas de séparation à un autre. L'écart entre les rapports source-à-interférences (SIR) obtenus en utilisant différentes valeurs du paramètre p peut s'expliquer par la convergence vers des minima locaux. Ceci peut être vérifié en remplaçant la méthode d'optimisation actuelle par une méthode qui peut éviter la convergence vers un minimum local, par exemple en utilisant le recuit simulé. Nous n'avons pas pu tester ce genre de méthodes mais ceci reste dans nos perspectives à court terme.

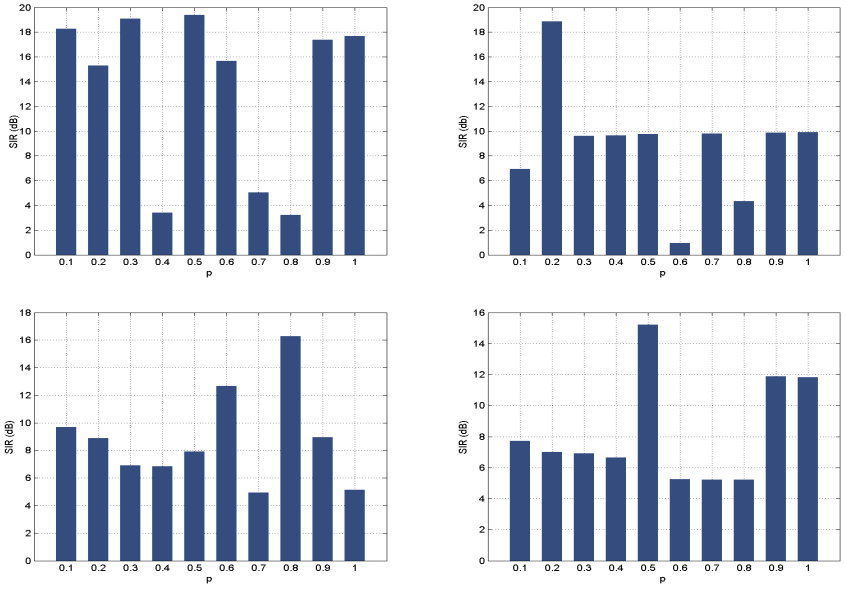


FIGURE 9.4 – Variation du rapport signal-à-interférences SIR de l'algorithme BSS-l_p en fonction du paramètre p : exemple de séparation de 4 paires de sources de la base de données Theo-RI-studio.

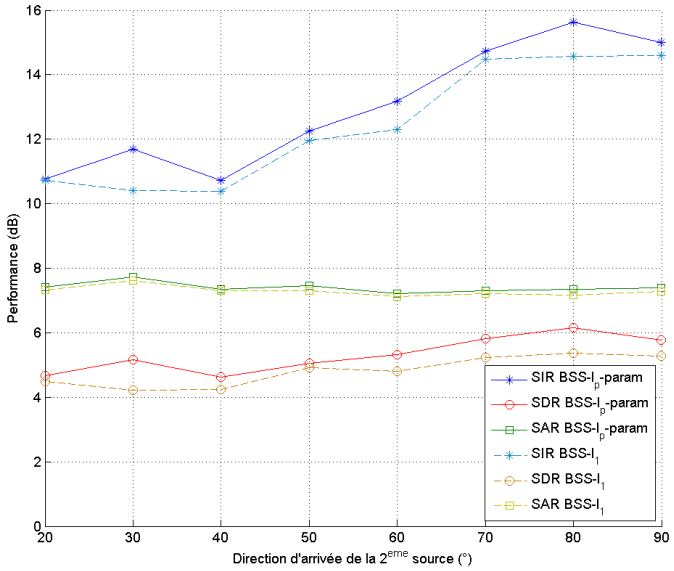


FIGURE 9.5 – Le rapport source-à-interférences SIR (courbes en bleu), le rapport source-à-distorsion SDR (courbes en rouge) et le rapport sources-à-artéfacts SAR (courbes en vert) de l'algorithme BSS-l_p-param (courbes continues) en comparaison avec BSS-l_1 (courbes en tirets) : évaluation sur la base de données Theo-RI-studio.

Nous avons donc essayé d'exploiter la norme l_p dans la séparation de source sous

une autre forme. Nous avons tenté de durcir la contrainte de parcimonie de la norme l_p au fur et à mesure que l'algorithme avance dans les itérations en faisant décroître le paramètre p selon une fonction sigmoïde comme nous l'avons présenté dans la section 4.4. Nous appelons cet algorithme BSS-l_p-param. La figure 9.5 montre le rapport source-à-interférences SIR, le rapport source-à-distorsion SDR et le rapport sources-à-artéfacts SAR de BSS-l_1 et BSS-l_p-param. Les performances de BSS-l_p-param sont légèrement supérieures à celles de l'algorithme BSS-l_1. L'analyse perceptuelle (*cf.* figure 9.6) montre que le score perceptuel relatif aux artéfacts de BSS-l_p-param est légèrement inférieur à celui de BSS-l_1. L'algorithme BSS-l_p-param a l'avantage d'avoir des écarts types inférieurs à ceux de BSS-l_1 pour les scores perceptuels et le rapport source-à-interférences (*cf.* figure 9.7).

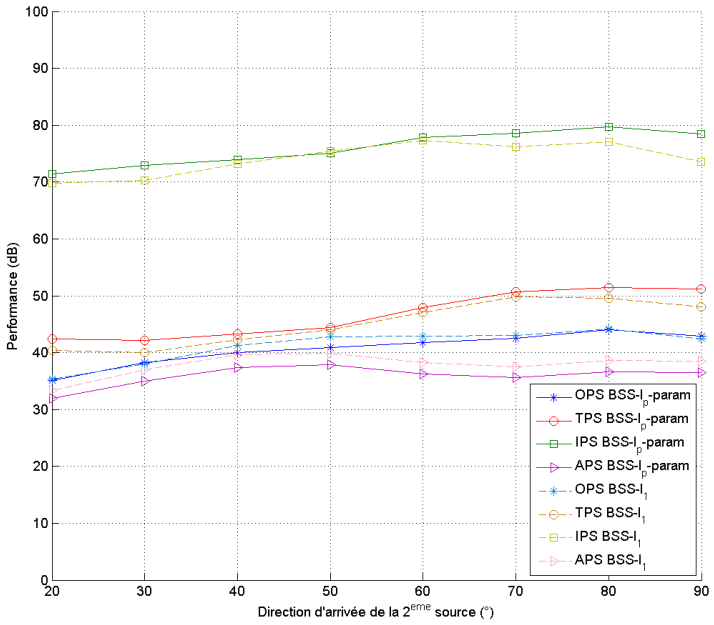


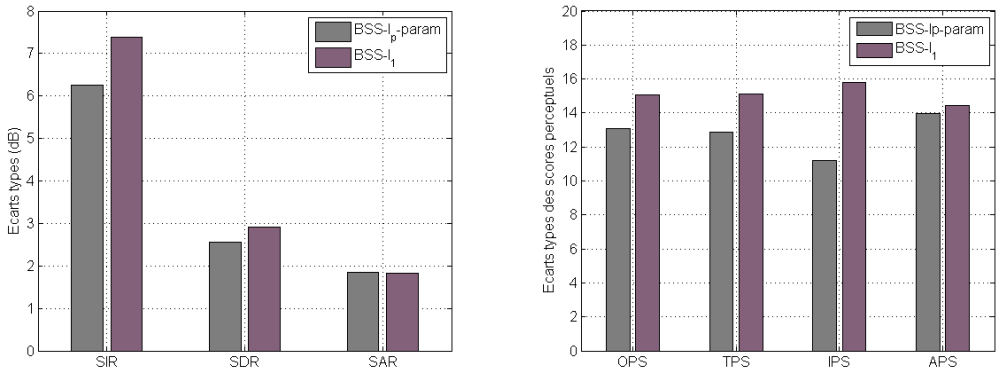
FIGURE 9.6 – Le score perceptuel global OPS (courbes en bleu), le score perceptuel relatif à la cible TPS (courbes en rouge), le score perceptuel relatif aux interférences IPS (courbes en vert) et le score perceptuel relatifs aux artéfacts APS (courbes en mauve) de l'algorithme BSS-l_p-param (courbes continues) en comparaison avec BSS-l_1 (courbes en tirets) : évaluation sur la base de données Theo-RI-studio.

La figure 9.8 montre les courbes de la parcimonie moyenne des sources estimées au cours des itérations mesurées par l'indice de Gini et pour les différentes directions d'arrivées considérées. L'indice de Gini est une bonne mesure de parcimonie pour la parole [69], il est compris entre 0 et 1 : plus le signal est parcimonieux, plus son indice de Gini est proche de 1. Nous remarquons qu'après une centaine d'itérations, ce qui correspond à un p inférieur à 0.5, les courbes de parcimonie décroissent légèrement.

Quand le paramètre p est inférieur à 0.5, nous remarquons que les performances de chaque paire de sources estimée sont chahutées (exemple du rapport source-à-interférences SIR de la figure 9.9 au cours des itérations et pour différentes directions d'arrivées), ce qui implique une baisse de la moyenne générale de la parcimonie.

Ceci peut être dû au fait que la contrainte de parcimonie devient très dure pour permettre une convergence stable de l'algorithme (si nous considérons le cas de chaque source estimée individuellement). Suite à cette constatation, deux pistes sont possibles : la première consiste à arrêter les itérations à 100 et donc à $p = 0.5$, la deuxième consiste à faire décroitre p de 1 à 0.5 selon le même type de sigmoïde en gardant le même nombre d'itérations. Ceci fera l'objet de nos futurs travaux. Dans la même figure 9.8, nous remarquons aussi que la parcimonie des sources estimées augmente à travers les itérations et converge rapidement vers la valeur optimale (moins de 30 itérations dans notre contexte).

L'algorithme de séparation de source avec la minimisation de la pseudo-norme l_p paramétrée est un algorithme prometteur qui ouvre la voie à l'étude de différentes configurations possibles de ce type de fonctions de coût paramétrée. En effet, plusieurs types de contraintes de parcimonie peuvent être étudiés en changeant la forme de la fonction logistique et la valeur minimale à atteindre du paramètre p. Une étude théorique de la convergence de ce genre d'algorithme est aussi nécessaire.



(a) Les écarts types des SIR, SDR et SAR (b) Les écarts types des OPS, TPS, IPS et ASP

FIGURE 9.7 – Les écarts types des résultats de séparations des algorithmes BSS-l_p-param (barres grises) et BSS-l_1 (barres violettes) : évaluation sur la base de données Theo-RI-studio.

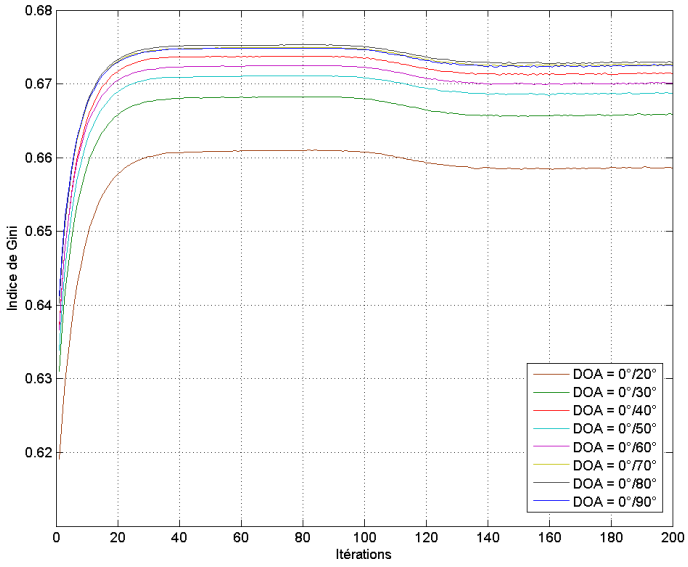


FIGURE 9.8 – L'indice de Gini moyen de BSS-l_p-param au cours des itérations sur la base de données Theo-IR-Studio.

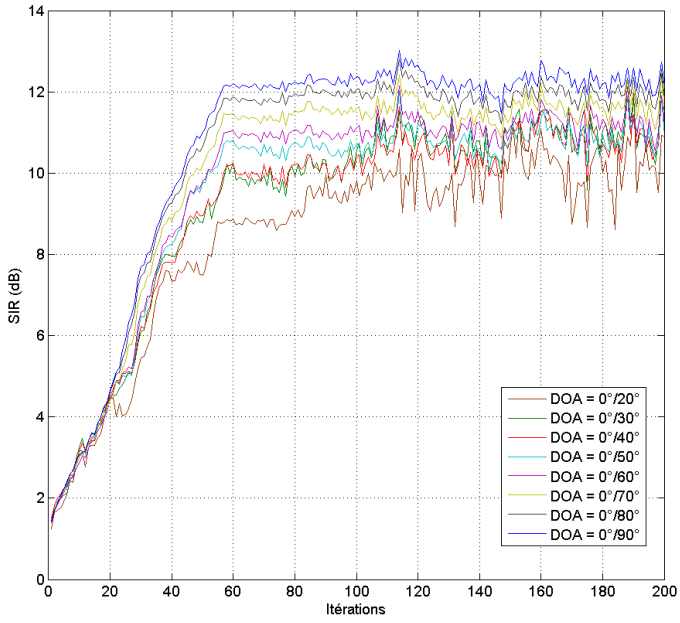


FIGURE 9.9 – Evaluation du rapport source-à-interférences SIR de l'algorithme BSS-l_p-param au cours des itérations sur la base de données Theo-RI-Studio pour les différentes directions d'arrivées DOA considérées.

9.3 Évaluation de l'apport du prétraitement par formation de voies des algorithmes de séparation à deux étapes

Nous étudions dans cette sous-section l'apport du prétraitement par formation de voies en termes de performance de séparation. Une partie de ces résultats a été publiée dans EURASIP Journal on Advances in Signal Processing (*cf.* [3, 4]). Nous considérons les configurations suivantes des algorithmes de séparation à deux étapes avec prétraitement de formation de voies :

- BF_DOA+BSS-l_1 : formation de N voies vers les directions d'arrivées suivi de BSS-l_1, l'algorithme de séparation de source avec la minimisation de la norme l_1 ;
- BF_fixed[30°]+BSS-l_1, BF_fixed[30°]+ICA : formation de 7 voies de -90° à 90° avec un angle inter-lobes égal à 30°, suivi respectivement de BSS-l_1 ou ICA ;
- BF_fixed[15°]+BSS-l_1 : formation de 13 voies de -90° à 90° avec un angle inter-lobes égal à 15° suivi de BSS-l_1 ;
- BF_fixed[10°]+BSS-l_1 : formation de 19 voies de -90° à 90° avec un angle inter-lobes égal à 10° suivi de BSS-l_1 ;
- BF_fixed[5°]+BSS-l_1 : formation de 37 voies de -90° à 90° avec un angle inter-lobes égal à 5° suivi de BSS-l_1 ;
- BF_fixed[30°]_BS+BSS-l_1 : formation de 7 voies de -90° à 90° avec un angle inter-lobes égal à 30° suivi d'une sélection de N lobes et BSS-l_1 ;
- BF_fixed[5°]_BS+BSS-l_1 : formation de 37 voies de -90° à 90° avec un angle inter-lobes égal à 5° suivi d'une sélection de N lobes et BSS-l_1.

9.3.1 Influence du prétraitement par formation de voies

Les figures 9.10 et 9.11 montrent que le rapport source-à-interférences SIR et le rapport source-à-distorsion SDR, obtenus sur la base de données Theo-RI-studio avec l'algorithme BF_fixed[5°]+BSS-l_1 sont supérieurs à ceux obtenus sans prétraitement par formation de voies avec l'algorithme BSS-l_1 et bien supérieur à ceux obtenus avec seulement l'étape de formation de voies BF_fixed[5°]. Nous prenons comme référence le SIR et le SDR des signaux reçus aux capteurs 1 et 2 de Theo que nous avons nommé signaux_capteurs dans les figures. Cependant, cette amélioration des performances de séparation apportée par le prétraitement par formation de voies fixe est limitée et n'atteint pas les performances apportées par le prétraitement par

formation de voies vers les directions d'arrivées BF_DOA+BSS-l_1 comme le montre les figures 9.10 et 9.11. Mais nous pouvons dépasser cette limitation par la sélection des lobes comme nous le verrons dans la suite. Les rapports sources-à-artéfacts SAR de l'algorithme de séparation par la minimisation de la norme l_1, de la formation de voies seule BF_fixed[5°], de l'algorithme de séparation par la minimisation de la norme l_1 avec formation de voies fixe BF_fixed[5°]+BSS-l_1 ou vers les directions d'arrivée des sources BF_DOA+BSS-l_1 sont proches.

La figure 9.14 montre que le prétraitement par formation de voies augmente le score perceptuel relatif à la source cible TPS, c'est à dire que les sources séparés sont mieux préservées avec le prétraitement par formation de voies. Cependant, cette amélioration des performances de séparation apportée par le prétraitement par formation de voies fixe ne concerne pas la suppression de la source cible. En effet, les scores perceptuels relatifs aux interférences IPS (*cf.* figure 9.15) des algorithmes BSS-l_1 et BF_fixed[5°]+BSS-l_1 sont proches alors que celui de l'algorithme avec la formation de voies vers les directions d'arrivée BF_DOA+BSS-l_1 est bien meilleur. D'après la figure 9.16, nous constatons que la formation de voies fixe avec un angle inter-lobes de 5° BF_fixed[5°]+BSS-l_1 introduit des artéfacts, ce qui se reflète aussi dans son score perceptuel global OPS (*cf.* figure 9.13). Ce résultat n'a pas été mis en évidence par l'analyse objective en termes de SIR, SDR et SAR.

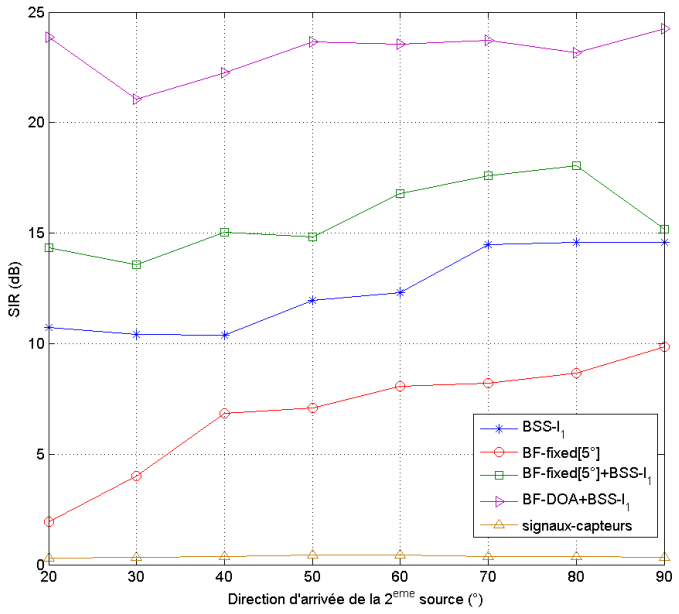


FIGURE 9.10 – Influence du prétraitement par formation de voies en termes de rapport source-à-interférences SIR : évaluation sur la base de données Theo-RI-studio.

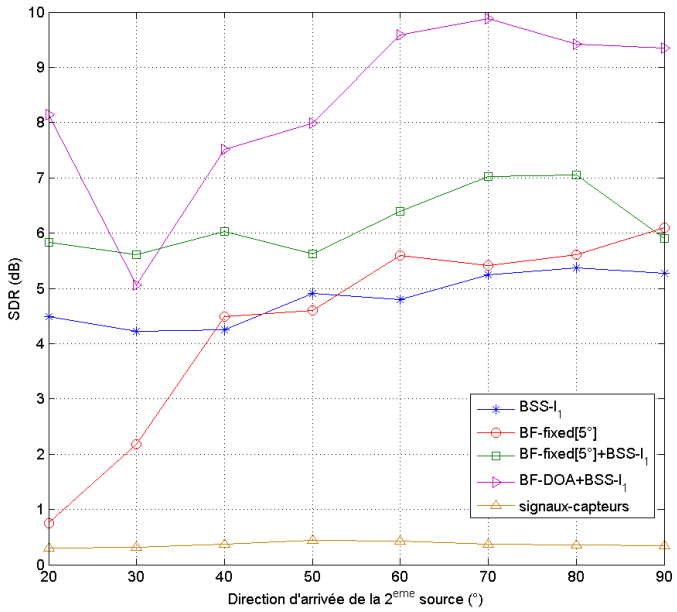


FIGURE 9.11 – Influence du prétraitement par formation de voies en termes de rapport source-à-distorsion SDR : évaluation sur la base de données Theo-RI-studio.

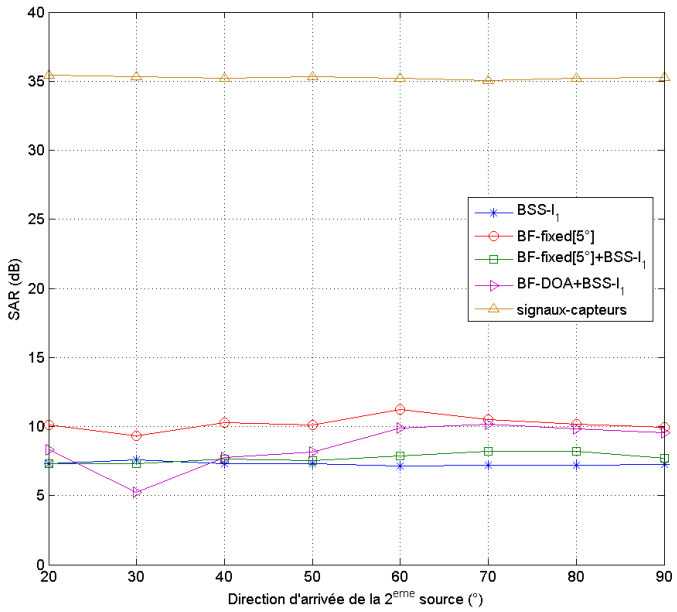


FIGURE 9.12 – Influence du prétraitement par formation de voies en termes de sources-à-artéfacts SAR : évaluation sur la base de données Theo-RI-studio.

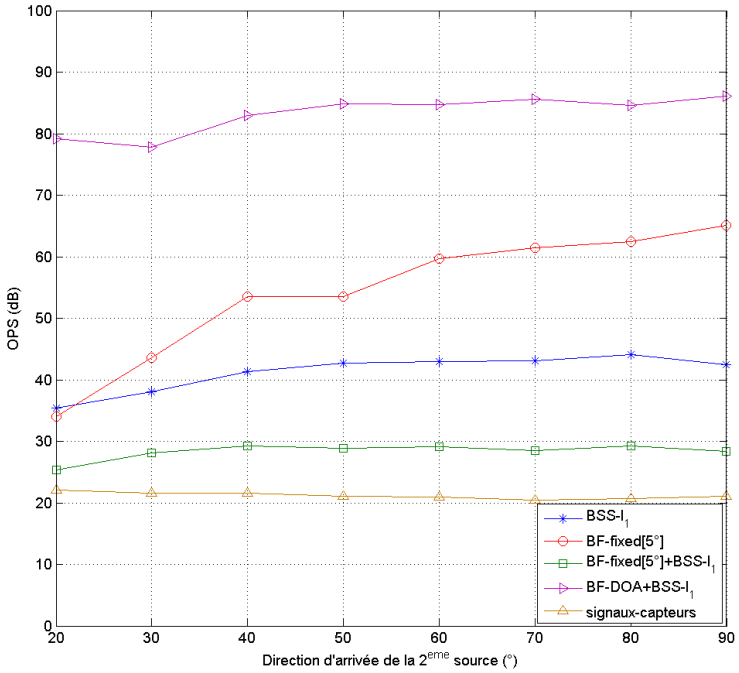


Figure 9.13 – Influence du prétraitement par formation de voies en termes de score perceptuel global OPS : évaluation sur la base de données Theo-RI-studio.

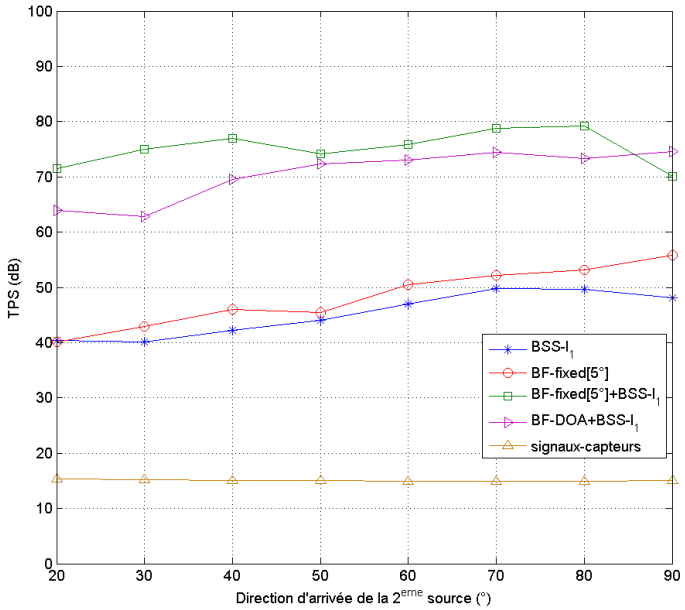


Figure 9.14 – Influence du prétraitement par formation de voies en termes de score perceptuel relatif à la source cible TPS : évaluation sur la base de données Theo-RI-studio.

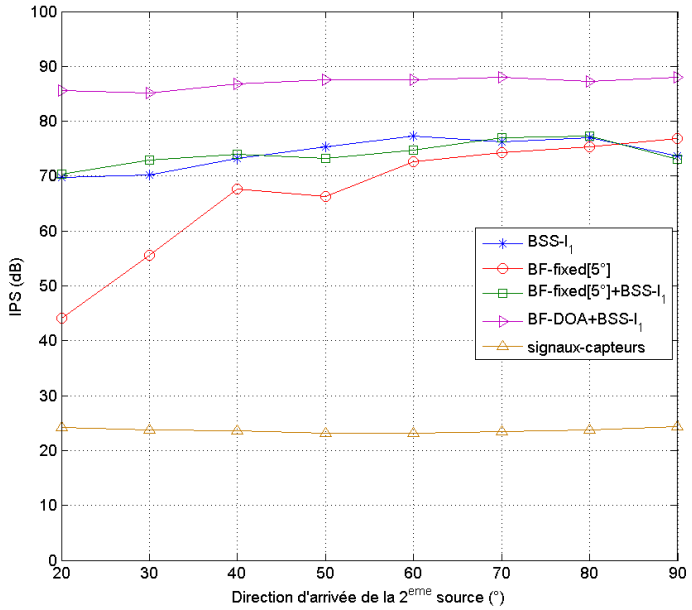


FIGURE 9.15 – Influence du prétraitement par formation de voies en termes de score perceptuel relatif aux interférences IPS : évaluation sur la base de données Theo-RI-studio.

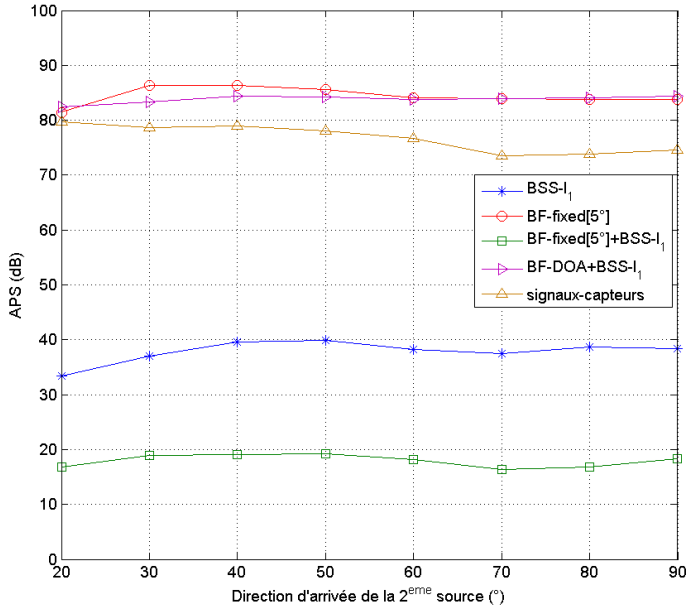


FIGURE 9.16 – Influence du prétraitement par formation de voies en termes de score perceptuel relatifs aux artéfacts APS : évaluation sur la base de données Theo-RI-studio.

9.3.2 Influence de la séparation angulaire entre les lobes :

Les figures 9.17, 9.18 et 9.19 montrent le rapport source-à-interférences SIR, le rapport source-à-distorsion SDR et le rapport sources-à-artéfacts SAR obtenus avec différents angles séparant les lobes de la formation de voies, ces configurations ont été présentées dans l'introduction de cette section. Les résultats montrent que quand nous augmentons le nombre de lobes dans l'espace de formation de voies qui nous intéresse, dans notre cas de -90° à 90°, le SIR et spécialement le SDR augmentent. En effet, en augmentant le nombre de lobes de la formation de voies, nous augmentons la chance de l'algorithme de tomber sur les bonnes directions d'arrivée des sources. Notons que dans les algorithmes BF_fixed[15°]+BSS-l_1 , BF_fixed[10°]+BSS-l_1 et BF_fixed[5°]+BSS-l_1, le prétraitement par formation de voies augmente le SDR par rapport à BSS-l_1. Le SIR des algorithmes avec prétraitement par formation de voies est meilleur que celui de l'algorithme de séparation seul BSS-l_1 et ceci dans toutes les configurations testées de la formation de voies fixes.

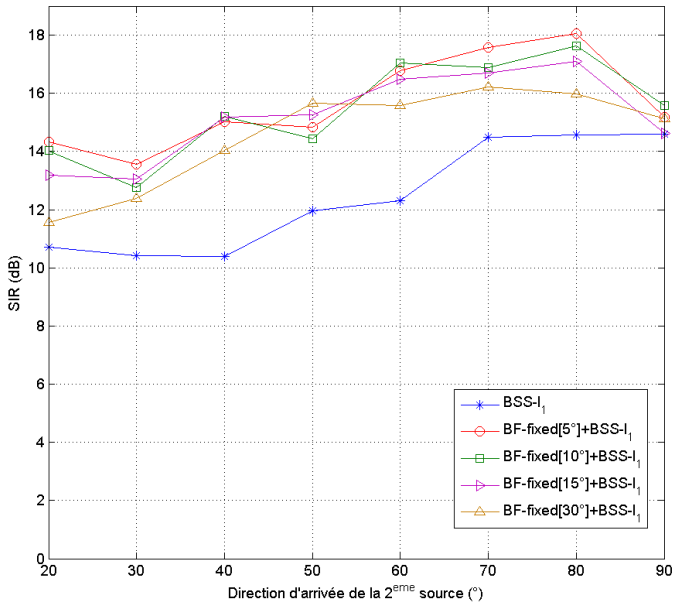


FIGURE 9.17 – Influence de l'angle inter-lobes de la formation de voies en termes de rapport source-à-interférences SIR : angle inter-lobes égal à 5° BF_fixed[5°]+BSS-l_1, 10° BF_fixed[10°]+BSS-l_1, 15° BF_fixed[15°]+BSS-l_1, et 30° BF_fixed[30°]+BSS-l_1, évaluation sur la base de données Theo-RI-studio.

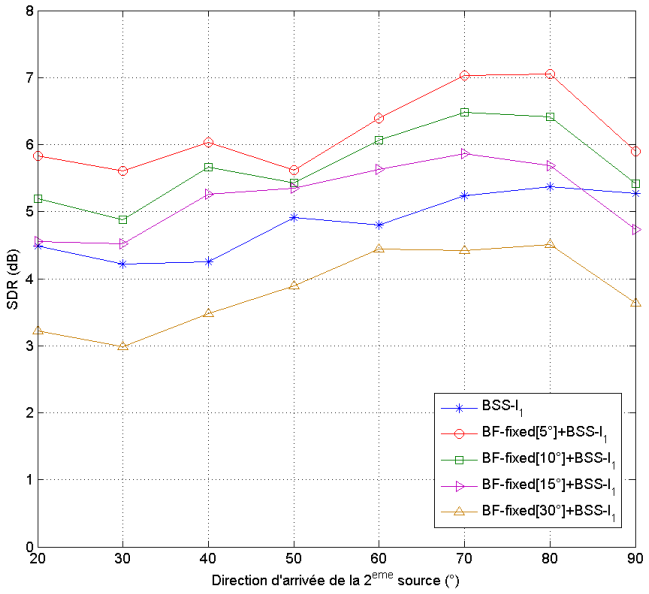


FIGURE 9.18 – Influence de l'angle inter-lobes de la formation de voies en termes de rapport source-à-distorsion SDR : angle inter-lobes égal à 5° BF_fixed[5°]+BSS-l_1, 10° BF_fixed[10°]+BSS-l_1, 15° BF_fixed[15°]+BSS-l_1, et 30° BF_fixed[30°]+BSS-l_1, évaluation sur la base de données Theo-RI-studio.

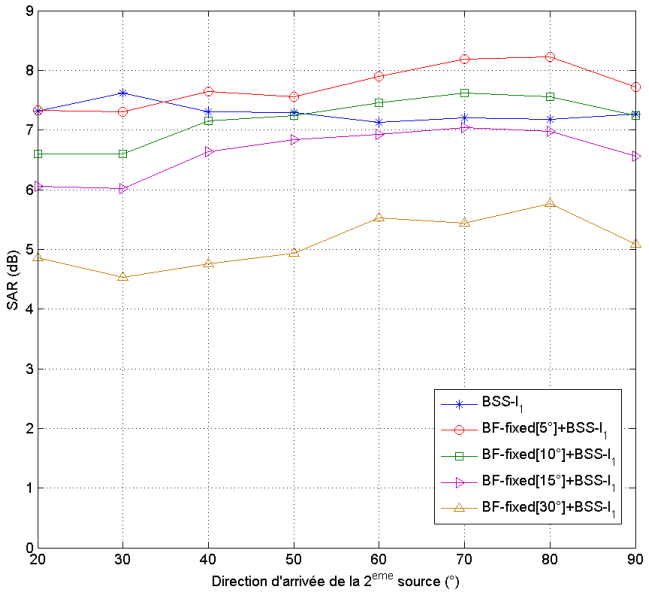


FIGURE 9.19 – Influence de l'angle inter-lobes de la formation de voies en termes de rapport sources-à-artéfacts SAR : angle inter-lobes égal à 5° BF_fixed[5°]+BSS-l_1, 10° BF_fixed[10°]+BSS-l_1, 15° BF_fixed[15°]+BSS-l_1, et 30° BF_fixed[30°]+BSS-l_1, évaluation sur la base de données Theo-RI-studio.

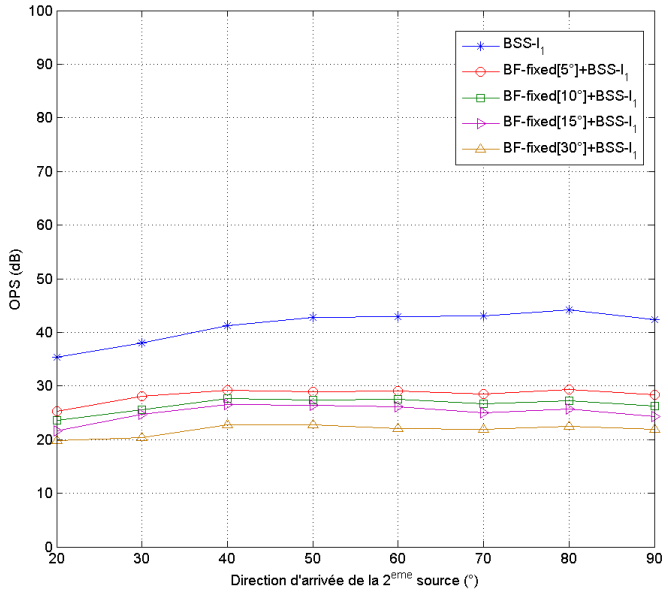


FIGURE 9.20 – Influence de l'angle inter-lobes de la formation de voies en termes de score perceptuel global OPS : angle inter-lobes égal à 5° BF_fixed[5°]+BSS-l_1, 10° BF_fixed[10°]+BSS-l_1, 15° BF_fixed[15°]+BSS-l_1, et 30° BF_fixed[30°]+BSS-l_1, évaluation sur la base de données Theo-RI-studio.

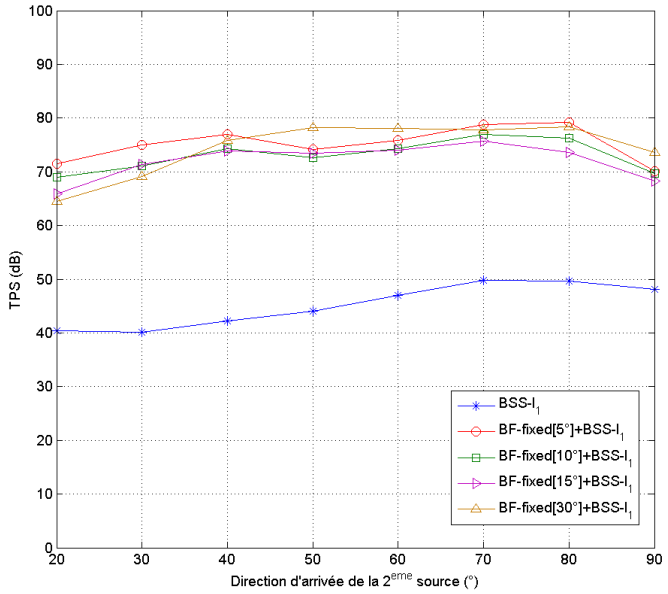


FIGURE 9.21 – Influence de l'angle inter-lobes de la formation de voies en termes de score perceptuel relatif à la cible TPS : angle inter-lobes égal à 5° BF_fixed[5°]+BSS-l_1, 10° BF_fixed[10°]+BSS-l_1, 15° BF_fixed[15°]+BSS-l_1, et 30° BF_fixed[30°]+BSS-l_1, évaluation sur la base de données Theo-RI-studio.

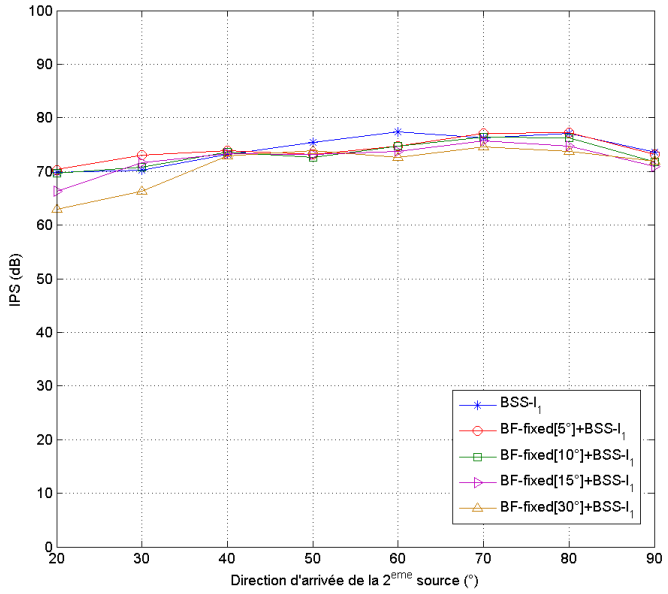


FIGURE 9.22 – Influence de l'angle inter-lobes de la formation de voies en termes de score perceptuel relatif aux interférences IPS : angle inter-lobes égal à 5° BF_fixed[5°]+BSS-l_1, 10° BF_fixed[10°]+BSS-l_1, 15° BF_fixed[15°]+BSS-l_1, et 30° BF_fixed[30°]+BSS-l_1, évaluation sur la base de données Theo-RI-studio.

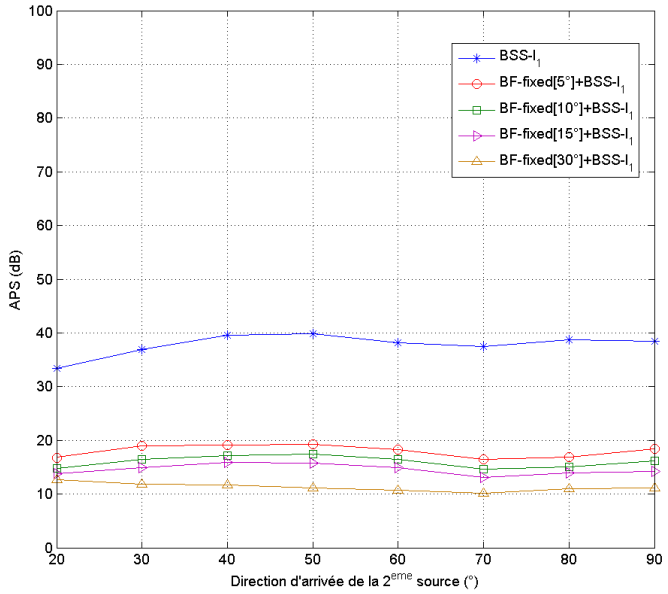


FIGURE 9.23 – Influence de l'angle inter-lobes de la formation de voies en termes de score perceptuel relatifs aux artéfacts APS : angle inter-lobes égal à 5° BF_fixed[5°]+BSS-l_1, 10° BF_fixed[10°]+BSS-l_1, 15° BF_fixed[15°]+BSS-l_1, et 30° BF_fixed[30°]+BSS-l_1, évaluation sur la base de données Theo-RI-studio.

Quant aux résultats perceptuels, ils confirment ce que nous avons constaté avec le prétraitement par formation de voies vers des directions fixes. Le prétraitement par formation de voies fixes, quel que soit l'angle entre les lobes, augmente le score perceptuel relatif à la source cible IPS par rapport à l'utilisation de BSS-l_1 seul (*cf.* figure 9.22). Par contre, le prétraitement par formation de voies fixe introduit des artéfacts comme le montre la figure 9.23 des scores perceptuels relatifs aux artéfacts APS, ce qui conduit à un score perceptuel global inférieur à celui de BSS-l_1 (*cf.* figure 9.20). Le score perceptuel relatif aux interférences est quant à lui presque le même avec ou sans prétraitement par formation de voies et pour tous les angles inter-lobes considérés.

Conclusion sur le prétraitement par formation de voies vers des directions de visée fixe :

Pour conclure sur cette partie, nous pouvons affirmer que le prétraitement par formation de voies vers des directions de visée fixes améliore les performances de séparation des sources audio par rapport à l'utilisation d'un algorithme de séparation seul et ceci pour tous les angles inter-lobes testés. Ceci est vrai pour toutes les mesures de performances à part le score perceptuel relatif aux artéfacts (APS) qui baisse lors de ce type de prétraitement. Néanmoins, avec le prétraitement par formation de voies vers des directions de visée fixes, nous n'atteignons pas encore les performances de séparations observées avec le prétraitement par formation de voies vers les directions d'arrivées.

9.3.3 Influence de la sélection de lobes

Comme nous pouvons le noter des figures 9.24, 9.25 9.26 et 9.31a, le prétraitement par formation de voies fixe avec sélection de lobes (BF_fixed[30°]_BS+BSS-l_1 et BF_fixed[5°]_BS+BSS-l_1) et le prétraitement par formation de voies fixe vers les directions d'arrivées des sources (BF_DOA+BSS-l_1) ont des performances proches en termes de rapport source-à-interférences SIR, de rapport source-à-distorsion SDR et de rapport sources-à-artéfacts SAR. Cependant, dans un milieu réverbérant où les directions d'arrivées ne peuvent pas être estimées avec précision, le prétraitement avec formation de voies fixe et sélection de lobes peut être une bonne solution pour améliorer le SIR et le SDR des sources estimées, par comparaison avec l'utilisation d'un algorithme de séparation seul (BSS-l_1).

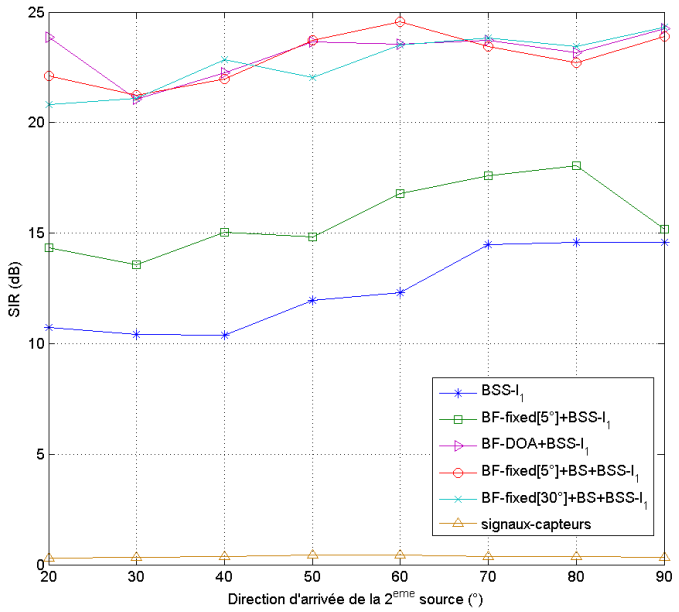


FIGURE 9.24 – Influence de la sélection de lobes en termes de rapport source-à-interférences SIR : évaluation sur la base de données Theo-RI-studio.

En comparant les algorithmes BF_fixed[30°]_BS+BSS-l_1 et BF_fixed[5°]_BS+BSS-l_1, nous pouvons noter que l'impact de l'angle inter-lobe est relativement faible par rapport au gain de performance et ceci grâce à la phase de sélection de lobes. Cependant, la formation de voies avec un angle inter-lobes de 5° nous permet d'estimer les directions d'arrivées des sources avec une résolution de 5° comme le montre la figure 9.32. Cette dernière représente une moyenne sur les lobes sélectionnés pour la base de données Theo-RI-studio, c'est à dire pour toutes les paires de sources à séparer et pour toutes les directions d'arrivées considérées.

L'analyse perceptuelle montre que la sélection de lobes améliore la qualité perceptuelle des sources séparées par rapport l'algorithme de séparation sans et avec prétraitement par formation de voies fixe. En effet, la figure 9.27 montre une amélioration du score perceptuel global OPS des sources séparées grâce à la phase de sélection de lobes. Plus particulièrement, le score perceptuel relatif à la source cible TPS (*cf.* figure 9.28), le score perceptuel relatif aux interférences IPS (*cf.* figure 9.29) et le score relatif aux artéfacts APS (*cf.* figure 9.30) ont augmenté. La sélection de lobes a aussi permis de réduire les écarts types des scores perceptuels TPS IPS et APS par rapport à ceux obtenus sans phase de sélection de lobes (*cf.* figure 9.31b).

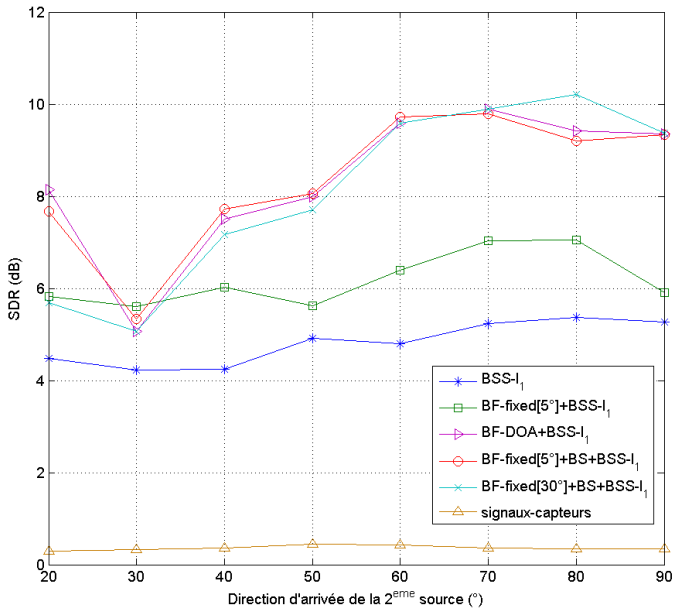


Figure 9.25 – Influence de la sélection de lobes en termes de rapport source-à-distorsion SDR : évaluation sur la base de données Theo-RI-studio.

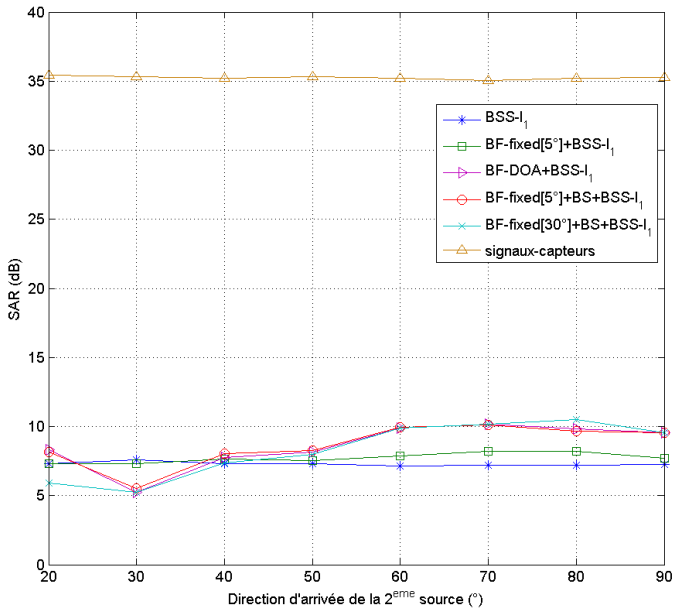


Figure 9.26 – Influence de la sélection de lobes en termes de rapport sources-à-artéfacts SAR : évaluation sur la base de données Theo-RI-studio.

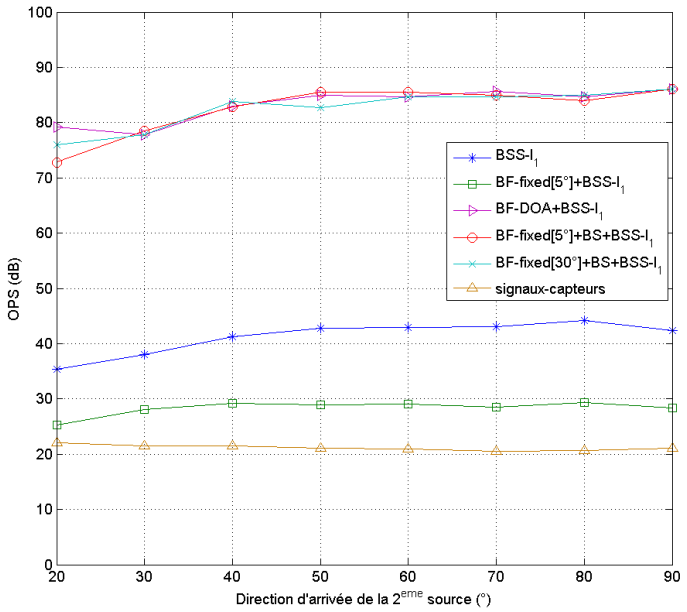


FIGURE 9.27 – Influence de la sélection de lobes en termes de score perceptuel global OPS : évaluation sur la base de données Theo-RI-studio.

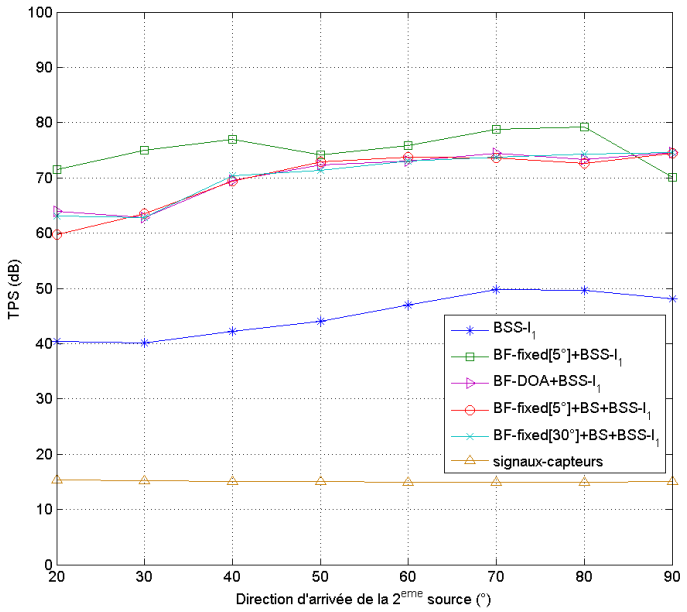


FIGURE 9.28 – Influence de la sélection de lobes en termes de score perceptuel relatif à la cible TPS : évaluation sur la base de données Theo-RI-studio.

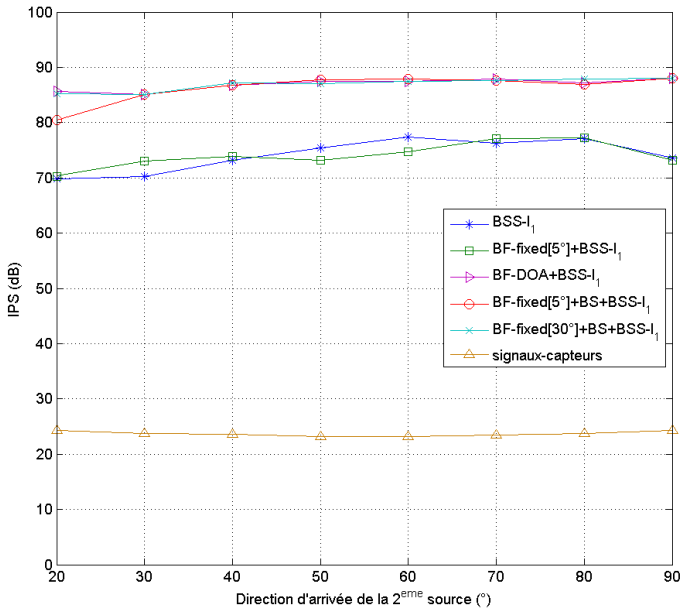


Figure 9.29 – Influence de la sélection de lobes en termes de score perceptuel relatif aux interférences IPS : évaluation sur la base de données Theo-RI-studio.

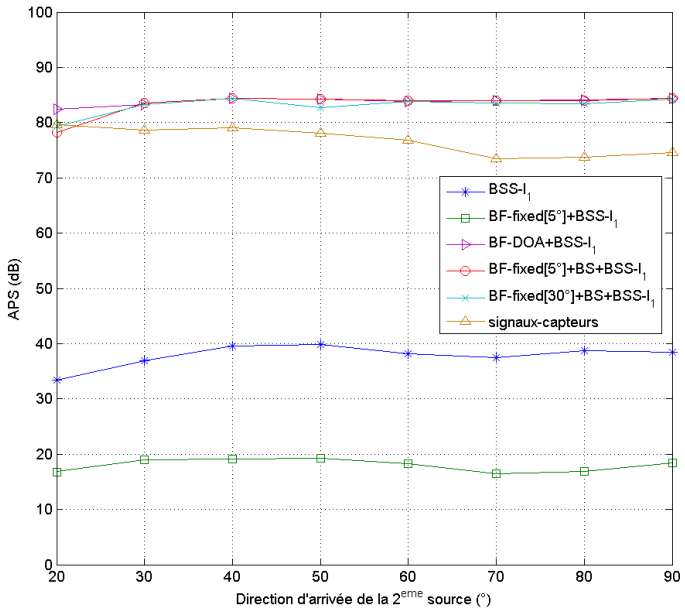
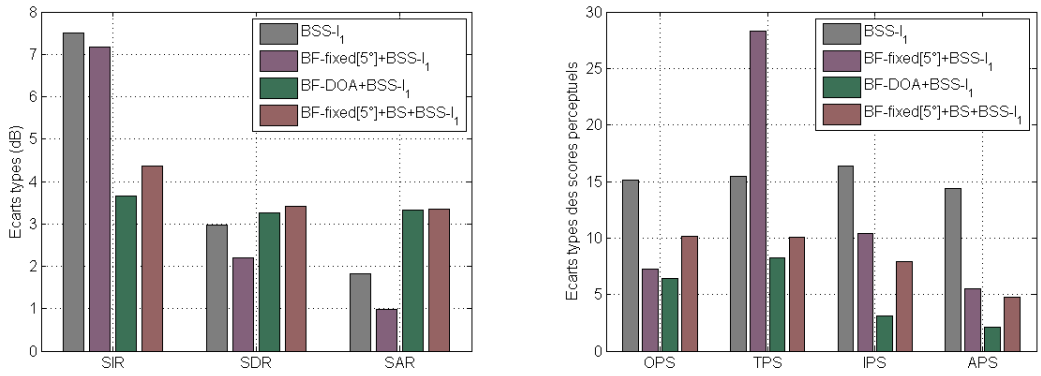


Figure 9.30 – Influence de la sélection de lobes en termes de score perceptuel relatifs aux artéfacts APS : évaluation sur la base de données Theo-RI-studio.



(a) Les écarts types des SIR, SDR et SAR (b) Les écarts types des OPS, TPS, IPS et ASP

FIGURE 9.31 – Les écarts types des résultats de séparations des algorithmes BSS-l_1 (barres grises), BF_fixed[5°]+BSS-l_1 (barres violettes), BF_DOA+BSS-l_1 (barres vertes) et BF_fixed[5°]_BS+BSS-l_1 (barres rouges) : évaluation sur la base de données Theo-RI-studio.

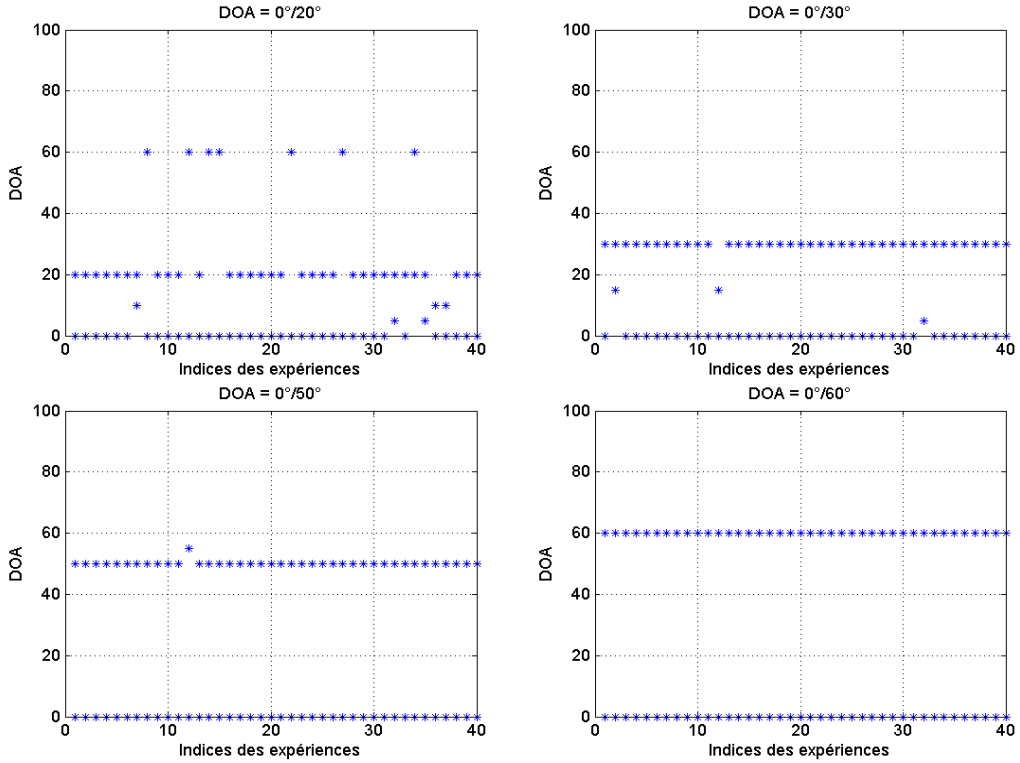
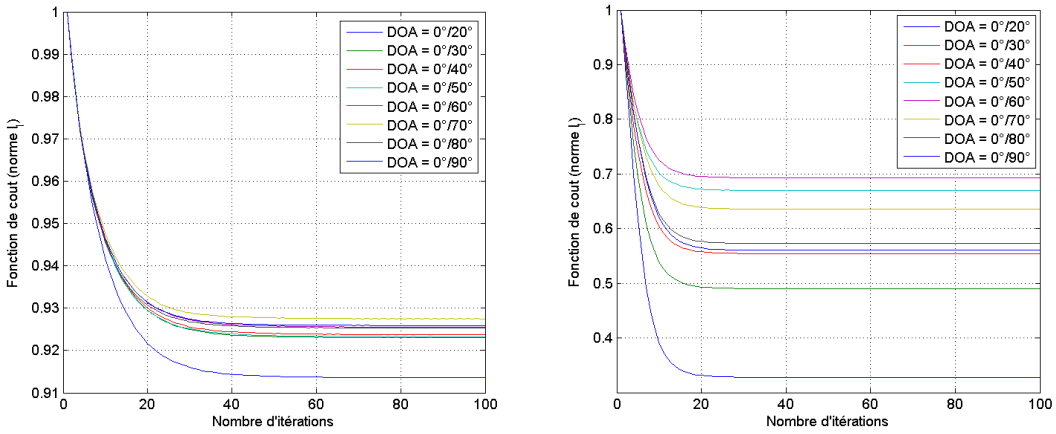


FIGURE 9.32 – Estimation des directions d'arrivées en utilisant l'algorithme de formation de voies BF_fixed[5°]_BS pour les 40 cas de séparation de source de la base de données Theo-RI-studio.

Les performances perceptuelles des algorithmes de séparation de sources avec prétraitement par formation de voies fixe et sélection de lobes (BF_fixed[30°]_BS+BSS-l_1 et BF_fixed[5°]_BS+BSS-l_1) restent proche de celles obtenues par l'algorithme avec prétraitement par formation de voies vers les directions d'arrivées des sources BF_DOA+BSS-l_1.

9.3.4 Analyse de la convergence

Nous avons procédé à l'analyse de la convergence de l'algorithme de séparation de sources avec prétraitement de formation de voies fixe et sélection de lobes BF_fixed[5°]_BS+BSS-l_1 en observant sa vitesse de convergence à travers les itérations pour les directions d'arrivées considérées (*cf.* figure 9.33). Chaque courbe représente la moyenne sur toutes les fréquences de la fonction de coût normalisée $\psi(f) = \frac{\sum_{i=1}^{N}\sum_{k=1}^{N_T}|Y_i(f,k)|}{\max\left(\sum_{i=1}^{N}\sum_{k=1}^{N_T}|Y_i(f,k)|\right)}$ (*cf.* section 4.3).



(a) (b)

FIGURE 9.33 – Étude de la vitesse de convergence : la valeur de la fonction de coût de (a) BSS-l_1 et (b) BF_fixed[5°]_BS+BSS-l_1 à travers les itérations et pour différentes directions d'arrivées.

Comme nous pouvons le voir dans la figure 9.33b, notre algorithme itératif à deux étapes converge assez rapidement (typiquement entre 10 et 20 itérations) vers un état stable. Nous notons aussi que la vitesse de convergence de la méthode avec prétraitement par formation de voies et sélection de lobes est meilleure que celle de BSS-l_1 seul (*cf.* figure 9.33a). En effet, dans ce contexte, l'algorithme de séparation

BSS-l_1 converge à son état stable après 30 à 40 itérations. De plus, la fonction de coût de l'algorithme à deux étapes atteint des valeurs plus basses que celles atteintes par BSS-l_1 et par conséquent, le traitement par formation de voies conduit à une meilleure convergence.

9.4 Variation des performances de séparation avec le nombre de capteurs

Nous nous intéressons ici à l'étude des performances de la séparation de sources audio par rapport au nombre de capteurs [6]. Cette étude a été effectuée avec notre algorithme de séparation à deux étapes BF_fixed[5°]_BS+BSS-l_1 en comparaison avec l'algorithme de la minimisation de la norme l_1 seul BSS-l_1. Par rapport à la méthode de séparation, notre but est d'étudier l'effet du nombre de capteurs sur la qualité de la séparation de sources. Plus spécifiquement, nous tentons de trouver le nombre optimal de capteurs qui doit être utilisé pour l'audition des robots, étant donnée une géométrie du réseau de capteurs. Nous montrons que dans notre cas, l'utilisation d'un réseau de capteurs augmente significativement les performances de séparation en comparaison avec le cas binaural (deux capteurs), cette augmentation se stabilise à partir d'un certain nombre de microphones que nous dévoilerons dans la suite.

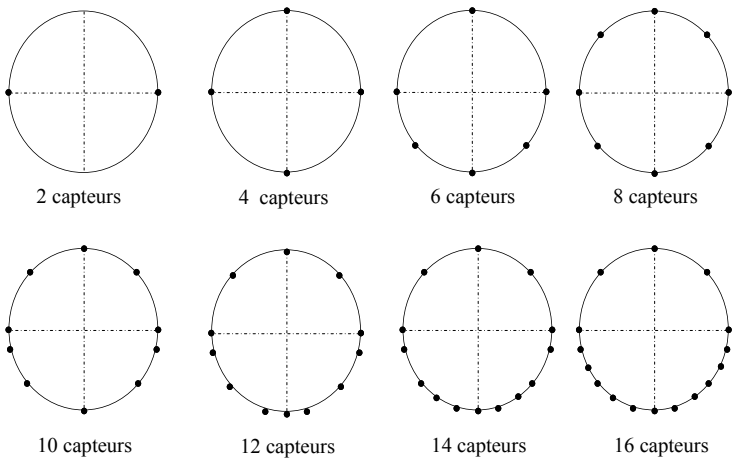


FIGURE 9.34 – Une vue de dessus des différentes configuration du réseau de capteurs : le nombre de microphones varie de 2 à 16.

Nous proposons d'évaluer le rapport source-à-interférences SIR, le rapport source-

à-distorsion SDR et le rapport sources-à-artéfacts SAR de la base de données Theo-RI-studio en variant le nombre de capteurs impliqué dans l'acquisition des mélanges de 2 à 16 et en considérant les configurations du réseau de microphones de la figure 9.34.

La figure 9.35 montre que, pour n'importe quel nombre de capteurs, la séparation aveugle de source avec un prétraitement par formation de voies et sélection de lobes BF_fixed[5°]_BS+BSS-l_1 a de meilleures performances que l'algorithme de séparation seul BSS-l_1. Les figures 9.35 et 9.36 montrent la variation des performances objectives et perceptuelles de l'algorithme BF_fixed[5°]_BS+BSS-l_1 avec le nombre de capteurs impliqué dans la séparation. Ces courbes représentent la moyenne des performances des cas de séparation de la base de données Theo-RI-studio. D'après ces courbes, les performances de séparation augmentent avec le nombre de capteurs.

La figure 9.37 montre la variation de la moyenne des SIR par rapport au nombre de capteurs. Nous pouvons noter une amélioration significative des performances quand le nombre de capteurs augmente. Quand le nombre de microphones M est supérieur ou égal à 8, il n'y a plus de gain significatif observé et ceci pour toutes les directions d'arrivées. Le même résultat est observé pour le SDR (*cf.* figure 9.38), SAR (*cf.* figure 9.39), et les scores perceptuels OPS (*cf.* figure 9.40), TPS (*cf.* figure 9.41), IPS (*cf.* figure 9.42) et APS (*cf.* figure 9.43). Le même phénomène se saturation des performances est observé pour la séparation de 3 et de 4 sources.

Ceci attire notre attention et peut être expliqué par : d'abord la réverbération modérée de la chambre dans laquelle les mesures ont été faites, nous nous attendons à un effet plus significatif de l'augmentation du nombre de capteurs dans un milieu plus réverbéré, ensuite il y a l'effet de la géométrie des lobes de la formation de voies. Nous notons aussi que plus la différence entre les directions d'arrivées est grande, plus la convergence par rapport au nombre de capteurs vers un état stable est rapide.

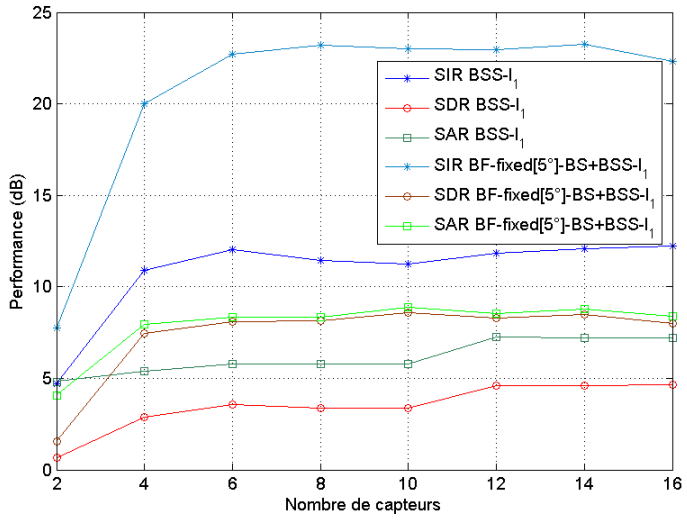


FIGURE 9.35 – Variation du rapport source-à-interférences SIR (courbes en bleu), le rapport source-à-distorsion SDR (courbes en rouge) et le rapport sources-à-artéfacts SAR (courbes en vert) avec le nombre de capteurs pour la séparation de deux sources à partir de différentes directions d'arrivées : évaluation de BF_fixed[5°]_BS+BSS-l_1 et BSS-l_1 : évaluation sur la base de données Theo-RI-studio.

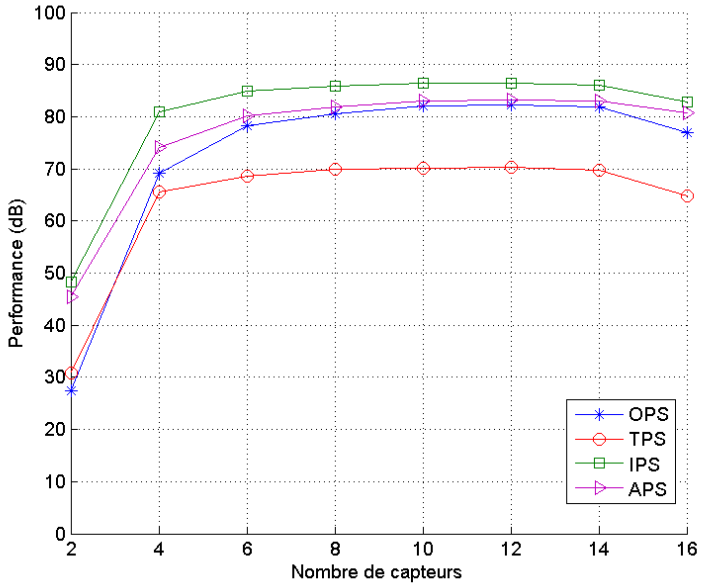


FIGURE 9.36 – Variation des mesures perceptuelles avec le nombre de capteurs pour la séparation de deux sources à partir de différentes directions d'arrivées : évaluation de BF_fixed[5°]_BS+BSS-l_1 : évaluation sur la base de données Theo-RI-studio.

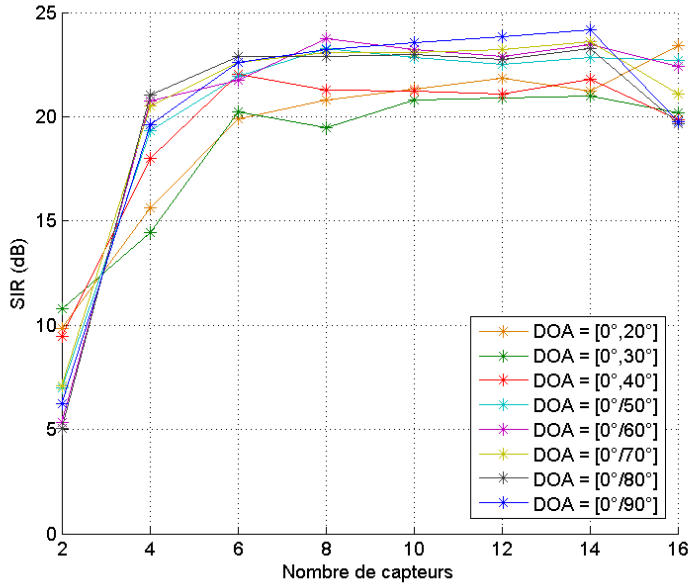


Figure 9.37 – Variation du rapport source-à-interférences SIR avec le nombre de capteurs et pour différentes directions d'arrivées pour l'algorithme BF_fixed[5°]_BS+BSS-l_1 : évaluation sur la base de données Theo-RI-studio.

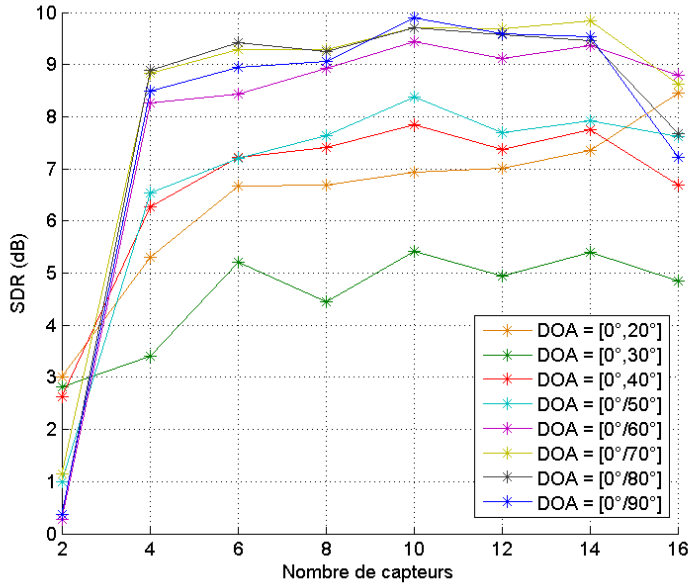


Figure 9.38 – Variation du rapport source-à-distorsion SDR avec le nombre de capteurs et pour différentes directions d'arrivées pour l'algorithme BF_fixed[5°]_BS+BSS-l_1 : évaluation sur la base de données Theo-RI-studio.

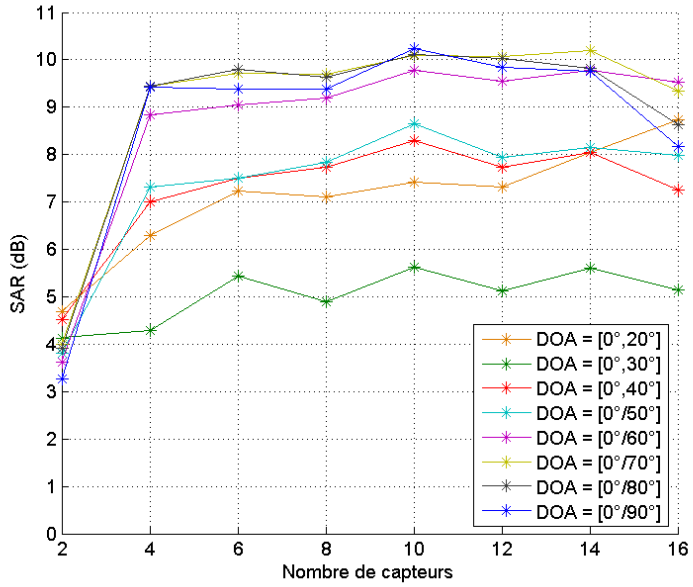


FIGURE 9.39 – Variation du rapport sources-à-artéfacts SAR avec le nombre de capteurs et pour différentes directions d'arrivées pour l'algorithme BF_fixed[5°]_BS+BSS-l_1 : évaluation sur la base de données Theo-RI-studio.

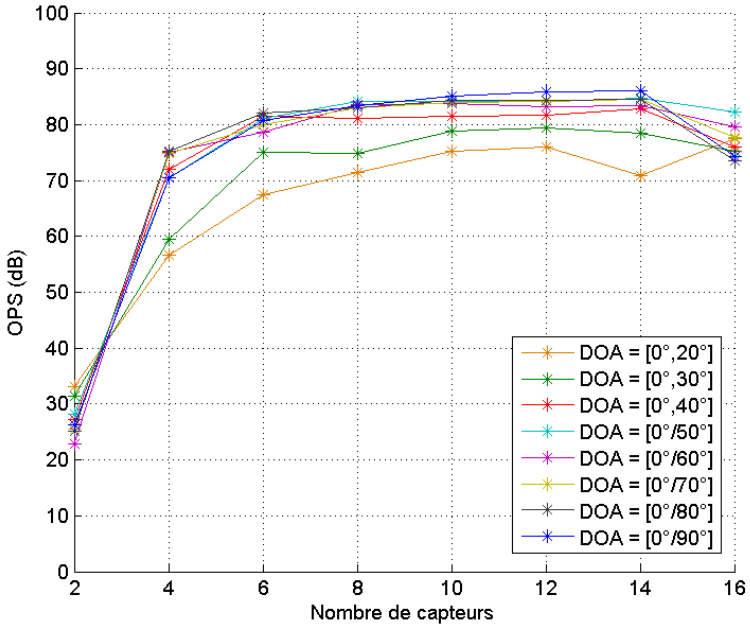


FIGURE 9.40 – Variation du score perceptuel global OPS avec le nombre de capteurs et pour différentes directions d'arrivées pour l'algorithme BF_fixed[5°]_BS+BSS-l_1 : évaluation sur la base de données Theo-RI-studio.

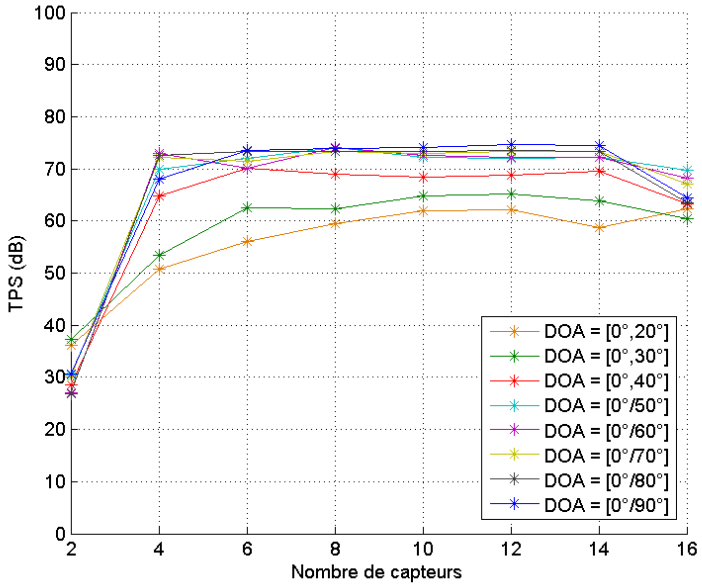


Figure 9.41 – Variation du score perceptuel relatif à la cible TPS avec le nombre de capteurs et pour différentes directions d'arrivées pour l'algorithme BF_fixed[5°]_BS+BSS-l_1 : évaluation sur la base de données Theo-RI-studio.

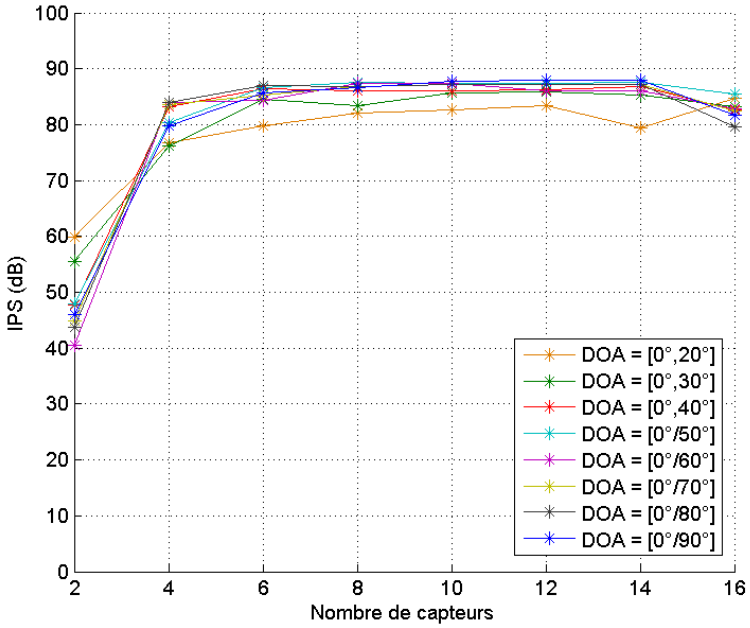


Figure 9.42 – Variation du score perceptuel relatif aux interférences IPS avec le nombre de capteurs et pour différentes directions d'arrivées pour l'algorithme BF_fixed[5°]_BS+BSS-l_1 : évaluation sur la base de données Theo-RI-studio.

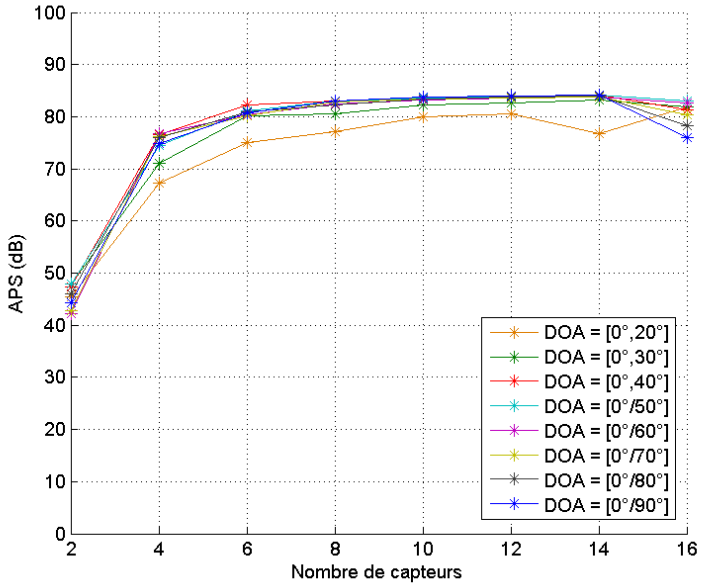


FIGURE 9.43 – Variation du score perceptuel relatif aux artéfacts APS avec le nombre de capteurs et pour différentes directions d'arrivées pour l'algorithme BF_fixed[5°]_BS+BSS-l_1 : évaluation sur la base de données Theo-RI-studio.

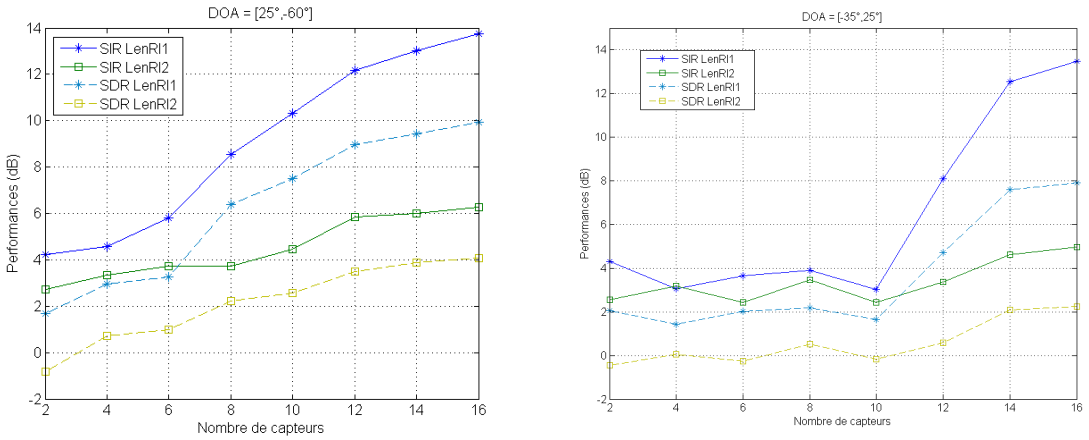


FIGURE 9.44 – Variation du rapport source-à interférences SIR et du rapport source-à-distorsion SDR de l'algorithme BF_fixed[5°]_BS+BSS-l_1 avec le nombre de capteurs : évaluation sur la base de données Theo-RI-IDV avec deux longueur de réponses impulsionnelles : RI1 = 500 échantillons et RI2 = 800 échantillons.

La figure 9.44 montre les courbes de rapport source-à-interférences SIR et du rapport source-à-distorsion SDR de BF_fixed[5°]_BS+BSS-l_1 en fonction de la variation du nombre de capteurs évaluées sur la base de données enregistrée à l'Institut de la Vision Theo-RI-IDV et pour deux différentes longueurs de réponses impulsionnelles :

RI1 = 500 échantillons et RI2 = 800 échantillons. La figure montre que les performances de séparation en termes de SIR et SDR ne semblent pas être stationnaire après un certain nombre de capteurs mais continuent à s'améliorer. Ce résultat ne va pas dans le sens de ce que nous avons observé pour la base de données Theo-RI-studio enregistrée dans le studio de Télécom ParisTech. Ceci peut s'expliquer par le fait que Theo-RI-IDV soit plus réverbérante que Theo-RI-studio. Nous pouvons supposer que si nous augmentons encore plus le nombre de capteurs (plus que 16 capteurs), nous pourrons observer le même phénomène de « saturation » des résultats de séparation observés dans le cas d'un environnement moins réverbérant.

A la figure 9.45, nous évaluons la vitesse de convergence de l'algorithme de séparation de sources avec prétraitement par formation de voies par rapport à la variation du nombre de capteurs. Nous notons que cette convergence est pratiquement indépendante de la taille du réseau de capteurs.

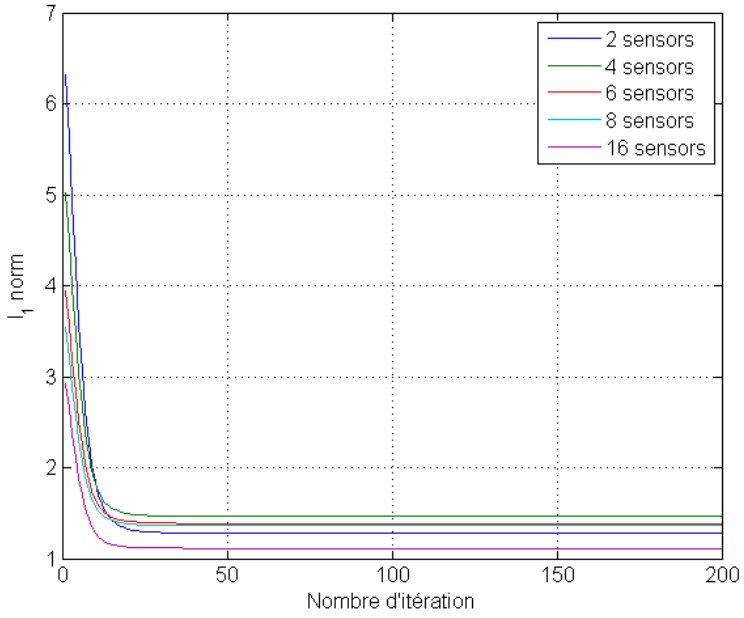


FIGURE 9.45 – La moyenne de la norme l_1 au cours des itérations de l'algorithme BF_fixed[5°]_BS+BSS-l_1 pour différentes tailles du réseau de capteurs : évaluation sur la base de données Theo-RI-studio.

Conclusion

Dans ce chapitre, nous avons présenté les résultats des algorithmes itératifs de séparation de sources que nous avons développés. Les principaux résultats de ce chapitre sont :

- la séparation de sources avec un critère de parcimonie basé sur la minimisation de la norme l_1 a d'aussi bonnes performances que l'analyse en composantes indépendantes basé sur la théorie de l'information ;
- la paramétrisation de la pseudo-norme l_p afin de rendre la contrainte de parcimonie de plus en plus dure au cours des itérations a des résultats prometteurs ;
- le prétraitement par formation de voies fixe utilisant les HRTFs améliore les performances moyenne de séparation (de 3dB maximum) ; en particulier le prétraitement par formation de voies fixes avec sélection de lobes améliore aussi bien la qualité objective (de 10 à 15 dB) que la qualité perceptuelle de la séparation (de 40 points environ) ;
- le nombre de capteurs nécessaire pour obtenir de bons résultats de séparation varie selon l'acoustique de l'environnement dans lequel se passe la séparation : pour la séparation d'un même mélange de sources, plus le milieu dans lequel nous effectuons cette séparation est réverbérant, plus le nombre de capteurs nécessaire pour obtenir une bonne séparation est grand.

Chapitre 10

Évaluation des algorithmes adaptatifs de séparation de sources

Introduction

Dans ce chapitre, nous évaluons les performances des algorithmes adaptatifs de séparation de sources présentés dans le chapitre 6. Nous évaluons nos algorithmes adaptatifs avec prétraitement de formation de voies sur des cas de séparation de deux sources où le nombre réel de sources peut être fixe ou variable. Nous proposons les trois scénarios d'évaluation suivant :

1. le nombre de sources à estimer est connu ; nous utilisons l'algorithme BF_fixed[5°]_BS+BSS-l_1 avec le vrai nombre de sources N ;
2. le nombre de sources à estimer est inconnu ; nous utilisons l'algorithme BF_fixed[5°]_BS+BSS-l_1 avec un nombre de sources à estimer fixe et égal à $N_{hyp} = 5$;
3. le nombre de source est estimé ; nous utilisons l'algorithme BF_fixed_NbSrEstim+BSS-l_1 qui nous permet d'estimer le nombre de sources dans l'étape de prétraitement par formation de voies.

Dans la suite, nous détaillerons les paramètres de ces algorithmes de séparation adaptatifs et nous procèderons à une évaluation de leurs performances en utilisant les outils d'évaluation présentés dans la section précédente : BSS_EVAL et PEASS.

Paramètres des algorithmes de séparation adaptatifs

Les signaux de test ont une durée de 5s pour le cas où le nombre réel de sources est fixe et 15s pour le cas où le nombre réel de sources est variable (*cf.* figure 10.1). Ces

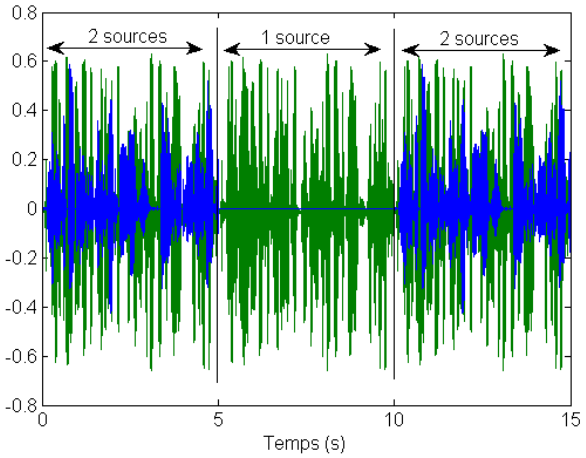


FIGURE 10.1 – Exemple d'un cas de séparation avant le mélange (sources bruts) avec un nombre de sources variant entre 1 et 2.

signaux sont échantillonnés à 16 kHz et sont extraits de la base de données Theo-RI-studio. Vingt cas de séparation ont été testés pour chacune des deux configurations. La fenêtre d'analyse spectrale est de type Hanning et de longueur 64 ms (1024 échantillons) et le pas d'avancement est de 50%. Le pas de mise à jour des algorithmes itératifs est $\mu = 0.05$. Nous choisissons un nombre d'itérations par fenêtre d'analyse temporelle longue égal à 2. Les résultats que nous présenterons dans la prochaine section sont la moyenne des résultats obtenus pour les 20 cas de séparation sur chaque fenêtre d'analyse longue.

10.1 Evaluation des algorithmes adaptatifs de séparation sans estimation du nombre de sources

10.1.1 Nombre de sources connu

Nous évaluons le rapport source-à-interférences SIR, le rapport source-à-distorsion SDR et le rapport sources-à-artéfacts SAR de l'algorithme adaptatif de séparation de sources avec un prétraitement par formation de voies et sélection de lobes BF_fixed[5°]_BS+BSS-l_1 pour deux configurations du nombre de sources : le nombre de sources est connu et fixe au cours du temps (*cf.* figures 10.2, 10.3 et 10.4) et le nombre de sources est connu et variable au cours du temps (*cf.* figures 10.5, 10.6 et 10.7). Le SIR et le SDR augmentent avec l'augmentation de l'angle séparant les deux sources et ceci pour toutes les itérations. Les rapports sources-à-artéfacts relatifs aux différentes directions d'arrivées quant à eux restent proches dans un intervalle de

5dB, comme observé dans le cas de la séparation itérative dans les figures 9.26 et 9.30. Pour ce cas de séparation de sources où le nombre de sources est connu et invariable au cours du temps, l'algorithme adaptatif BF_fixed[5°]_BS+BSS-l_1 présente de bonnes performances de séparation.

Dans le cas où le nombre de sources est variable, quand une seule source est active (typiquement entre les fenêtres d'analyses longues 157 et 299) et que cette source est bien extraite, le rapport source-à-interférences SIR estimé par BSS_EVAL est très élevé. Nous l'avons donc limité à 30dB dans la figure 10.5 pour des raisons pratiques. Nous remarquons que les performances de séparation baissent juste avant et après le changement du nombre de sources (*cf.* figures 10.5, 10.6 et 10.7). En effet, la source qui s'évanouit est de moins en moins présente dans la fenêtre d'analyse glissante, et quand seulement quelques échantillons relatifs à cette source sont présents dans cette fenêtre, l'algorithme de séparation a des difficultés à l'extraire.

Pour ce cas de séparation de sources où le nombre de sources est connu et variable au cours du temps, l'algorithme adaptatif BF_fixed[5°]_BS+BSS-l_1 suit bien le changement du nombre de sources et donne de bonnes performances de séparation.

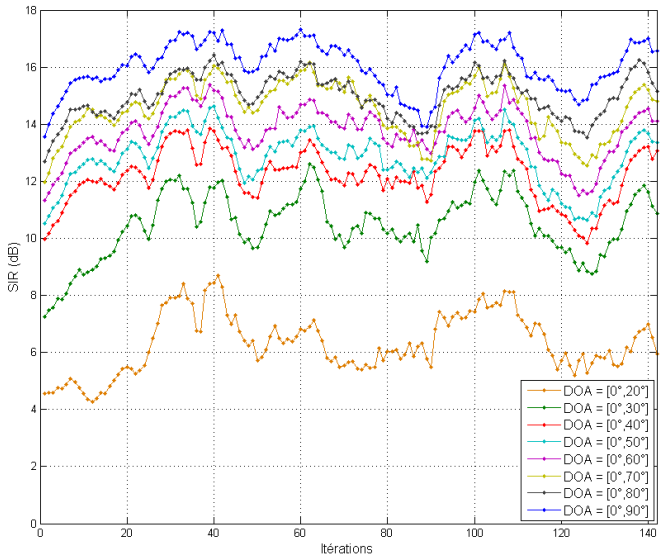


FIGURE 10.2 – Le rapport source-à-interférences SIR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse longues : nombre de sources réel fixe et connu (2 sources).

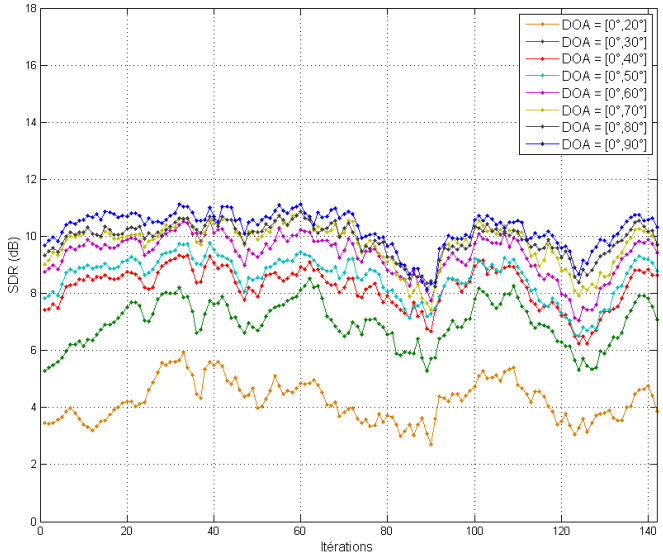


FIGURE 10.3 – Le rapport source-à-distorsion SDR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse longue : nombre de sources réel fixe et connu (2 sources).

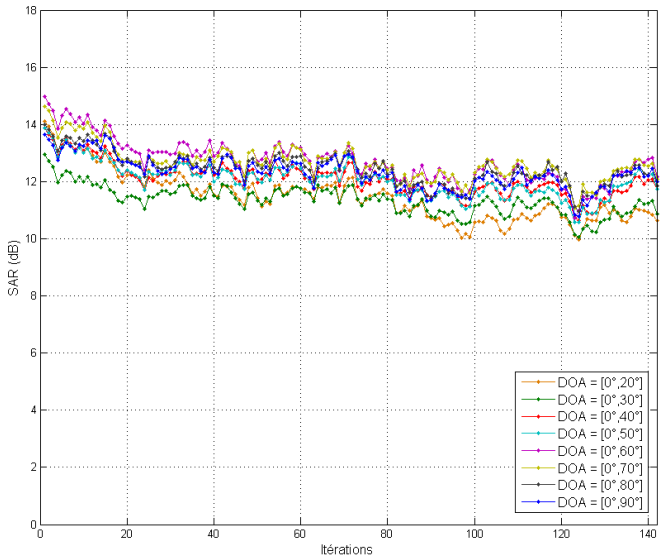


FIGURE 10.4 – Le rapport sources-à-artéfacts SAR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse longues : nombre de sources réel fixe et connu (2 sources).

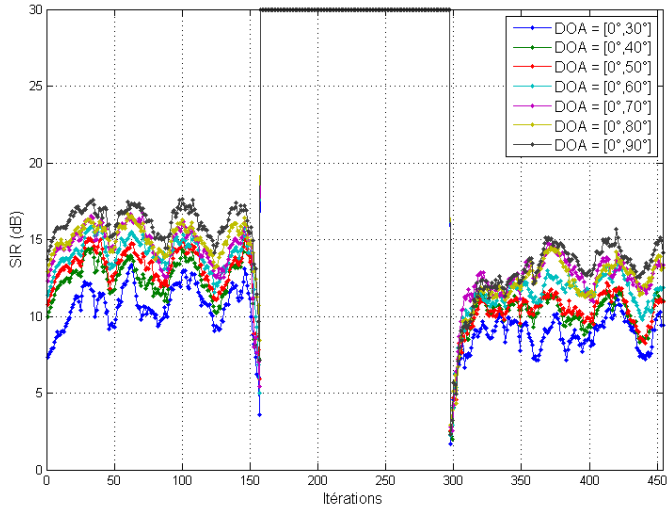


FIGURE 10.5 – Le rapport source-à-interférences SIR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse longues : nombre de sources réel connu et variable entre 1 et 2 sources.

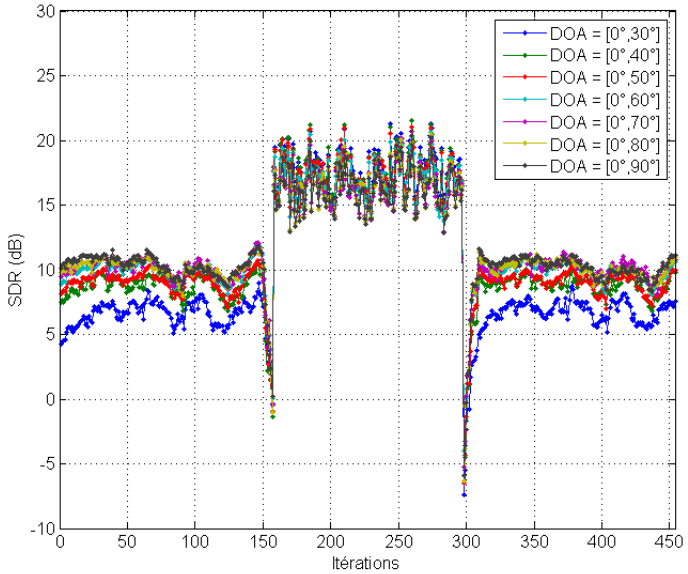


FIGURE 10.6 – Le rapport source-à-distorsion SDR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse longues : nombre de sources réel connu et variable entre 1 et 2 sources.

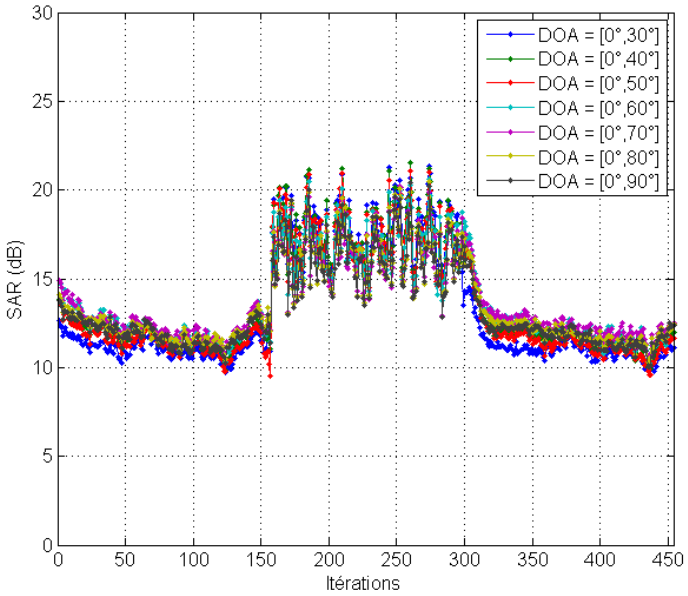


Figure 10.7 – Le rapport source-à-artéfacts SAR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse longues : nombre de sources réel connu et variable entre 1 et 2 sources.

10.1.2 Nombre de sources fixé *a priori*

Nous évaluons ici le SIR, le SDR et le SAR de l'algorithme de séparation adaptatif avec prétraitement par formation de voies BF_fixed[5°]_BS+BSS-l_1 pour le cas où le nombre de sources est fixe (*cf.* figures 10.8, 10.9 et 10.10) ou variable (*cf.* figures 10.11, 10.12 et 10.13). Pour chacune de ces deux configurations, nous séparons un nombre de sources fixe $N_{hyp} = 5$, supérieur au nombre de sources réel.

Que ce soit pour le cas où le nombre de sources est fixe (*cf.* figures 10.8, 10.9 et 10.10) ou variable (*cf.* figures 10.11, 10.12 et 10.13), les performances de la séparation sont proches de celles obtenues en connaissant le nombre réel de sources.

L'avantage de cette méthode de séparation adaptative est qu'elle ne nécessite pas une estimation du nombre de sources actives. Cependant, le nombre de sources à extraire doit être fixé de telle sorte qu'il soit supérieur au vrai nombre de sources. Ceci suppose une certaine connaissance de l'environnement dans lequel se déroule la séparation de sources et du nombre maximal de sources actives.

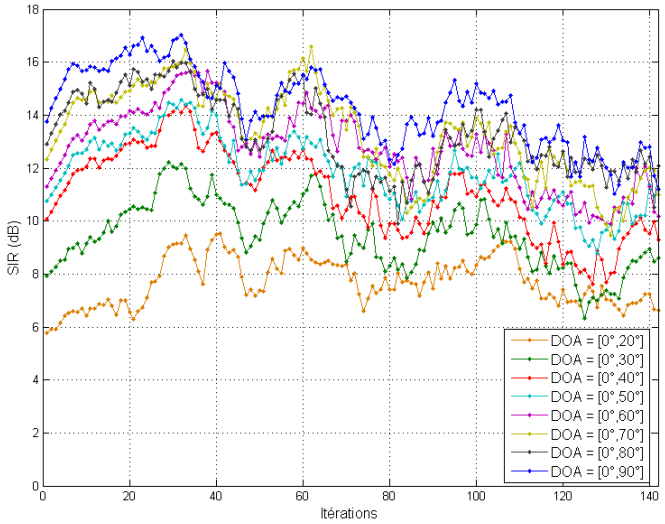


FIGURE 10.8 – Le rapport source-à-interférences SIR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse : nombre de sources réel est inconnu et fixé *a priori*.

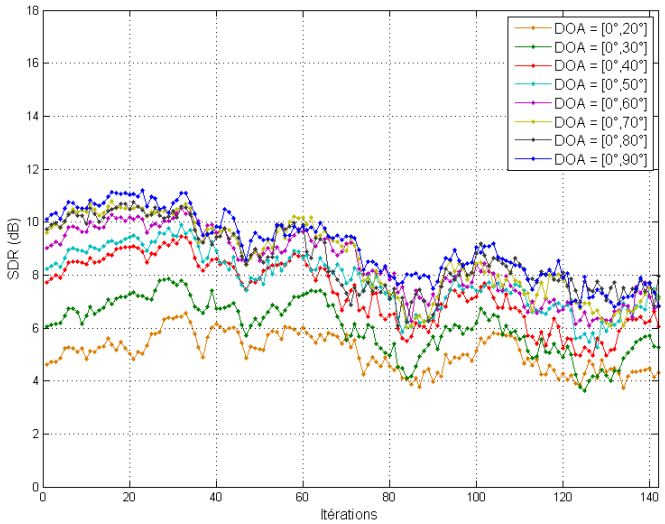


FIGURE 10.9 – Le rapport source-à-distorsion SDR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse : nombre de sources réel est inconnu et fixé *a priori*.

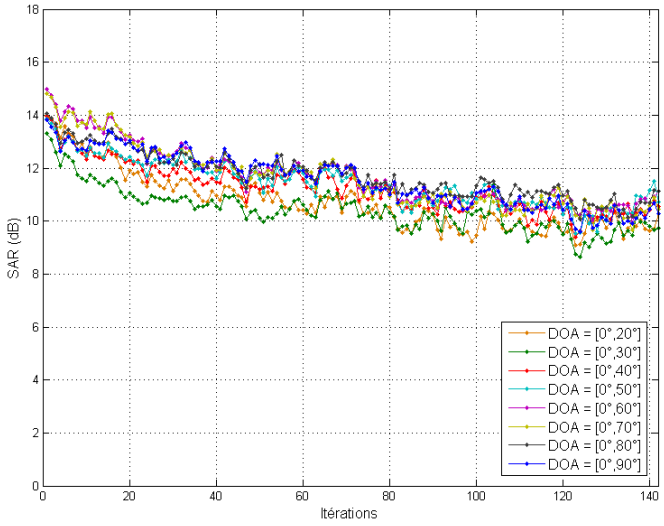


FIGURE 10.10 – Le rapport sources-à-artéfacts SAR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse : nombre de sources réel connu et variable entre 1 et 2 sources.

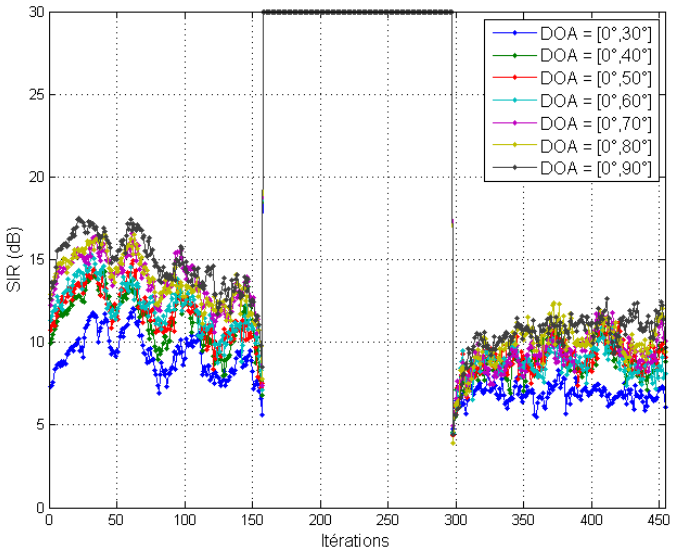


FIGURE 10.11 – Le rapport source-à-interférences SIR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse : le nombre de sources réel est inconnu et fixé *a priori*.

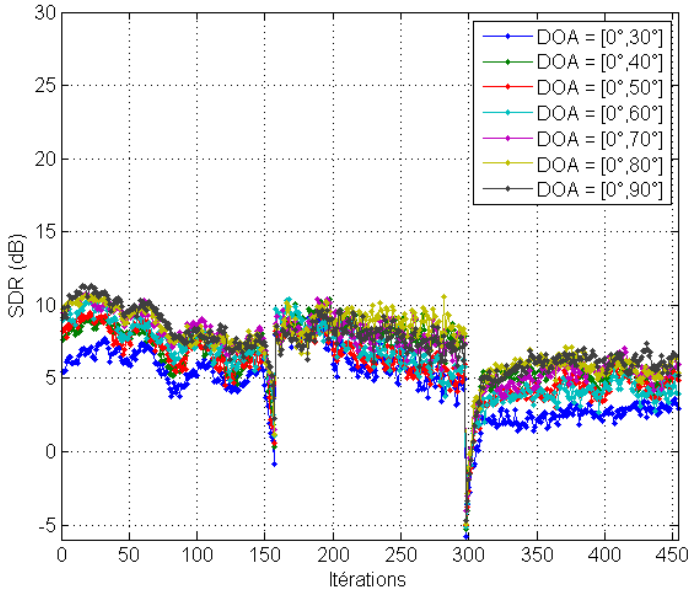


FIGURE 10.12 – Le rapport source-à-distorsion SDR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse : le nombre de sources réel est inconnu et fixé *a priori*.

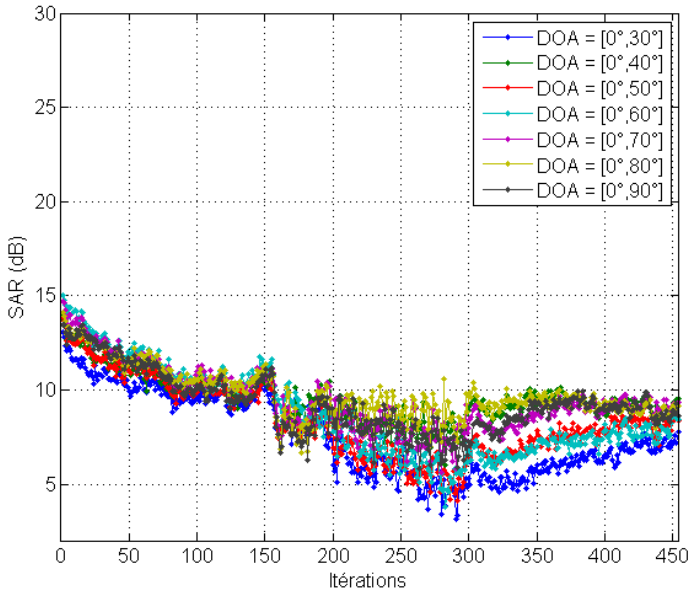


FIGURE 10.13 – Le rapport source-à-artéfacts SAR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse : le nombre de sources réel est inconnu et fixé *a priori*.

10.2 Evaluation des algorithmes de séparation avec estimation du nombre de sources

Dans cette section, nous évaluons notre algorithme de séparation en estimant le nombre de sources [2].

Les figures 10.14 et 10.15a montrent le nombre moyen de sources estimé au cours des fenêtres d'analyse longues où le nombre de sources à estimer est fixe et variable respectivement. Ces résultats ont été obtenus avec le prétraitement par formation de voies fixe et sélection des lobes plus grandes énergies en fixant l'angle inter-lobes à 5°. Nous comparons notre méthode d'estimation du nombre de sources à deux méthodes EIG1 [48] et EIG2 basée sur un simple seuillage des valeurs propres ordonnées des matrices de covariance des signaux reçus dans le domaine temps-fréquence. Nos résultats sont proches de EIG2 mais dans notre cas, l'estimation du nombre de sources est un résultat direct du prétraitement par formation de voies, elle est simple à implémenter et ne demande pas plus de calcul que celui de l'estimation de pics. La méthode EIG2 nécessite plus de temps de traitement que notre méthode à cause du calcul des matrices de covariances et de la décomposition en valeurs propres.

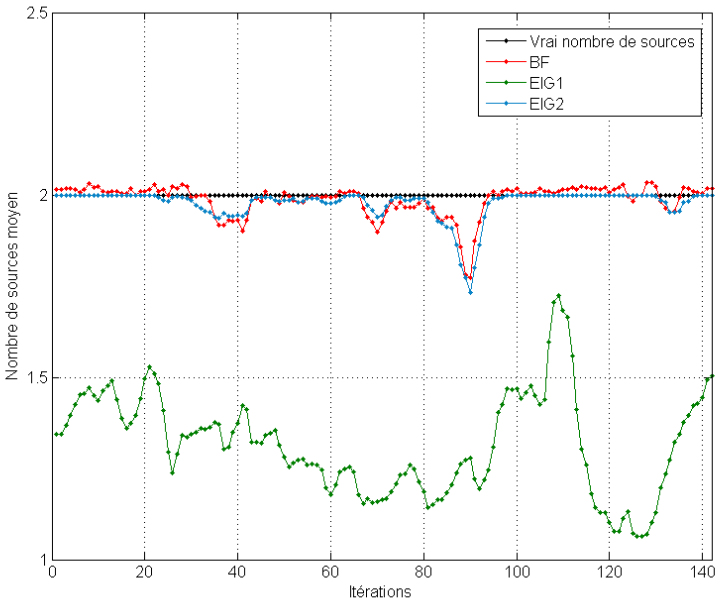
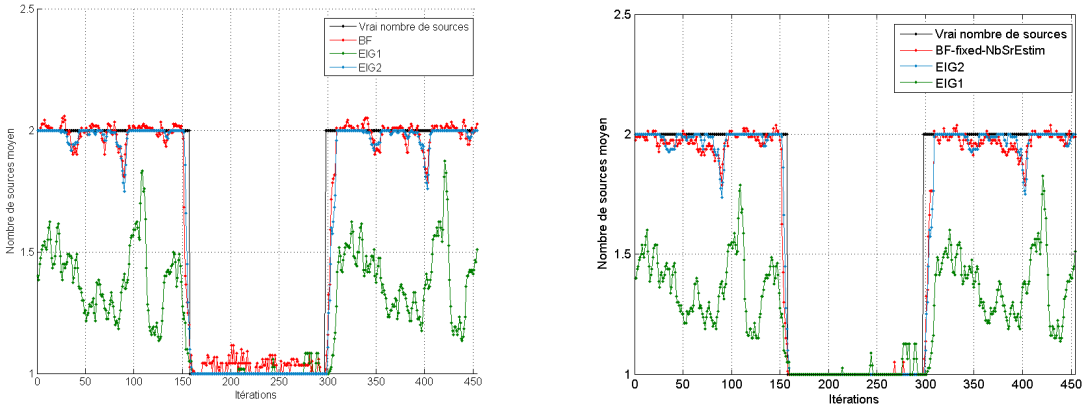


Figure 10.14 – Estimation du nombre de source au cours des fenêtres d'analyse avec BF_fixed[5°]_NbSrEstim : le nombre réel de sources est fixe.



(a) BF_fixed[5°]_NbSrEstim (b) BF_fixed[10°]_NbSrEstim

FIGURE 10.15 – Estimation du nombre de source au cours des fenêtres d'analyse avec BF_fixed_NbSrEstim : le nombre réel de sources varie entre 1 et 2.

Si dans le prétraitement par formation de voies fixe nous choisissons un angle inter-lobes égal à 10°, nous remarquons qu'il y a moins de sur-estimation du nombre de sources avec la méthode du prétraitement par formation de voies BF_fixed[10°]_NbSrEstim par comparaison à BF_fixed[10°]_NbSrEstim et ceci dans le cas où une seule sources est active (*cf.* figure 10.15b). Rappelons que pour l'estimation du nombre de sources, après le filtrage par formation de voies, nous sélectionnons N_{max} directions de visée qui correspondent aux N_{max} lobes ayant les plus grandes énergies. Ensuite, un histogramme qui correspond au nombre d'occurrence des directions d'arrivées sélectionnées est construit et le nombre de pics dont la valeur est supérieure à un seuil donné correspond au nombre de sources. La résolution angulaire de la formation de voies obtenue avec un angle inter-lobes égal à 5° est plus fine que celle obtenue avec un angle inter-lobes de 10°. Dans le cas où l'angle inter-lobes est égal à 5°, nous voyons donc apparaître des pics relatifs à la réverbération par exemple, ce qui peut conduire à une sur-estimation du nombre de sources actives. Cette sur-estimation du nombre de sources peut être évitée en utilisant le « bon » seuillage.

Les figures 10.16, 10.17 et 10.4 montrent le rapport source-à-interférences SIR, le rapport source-à- distorsion SDR et le rapport sources-à-artéfacts SAR moyens dans les cas de séparation où le nombre de sources ne change pas au cours du temps. Les performances de séparation de l'algorithme adaptatif avec l'estimation du nombre de sources dans ce cas restent proches de celles où le nombre de sources à séparer est connu.

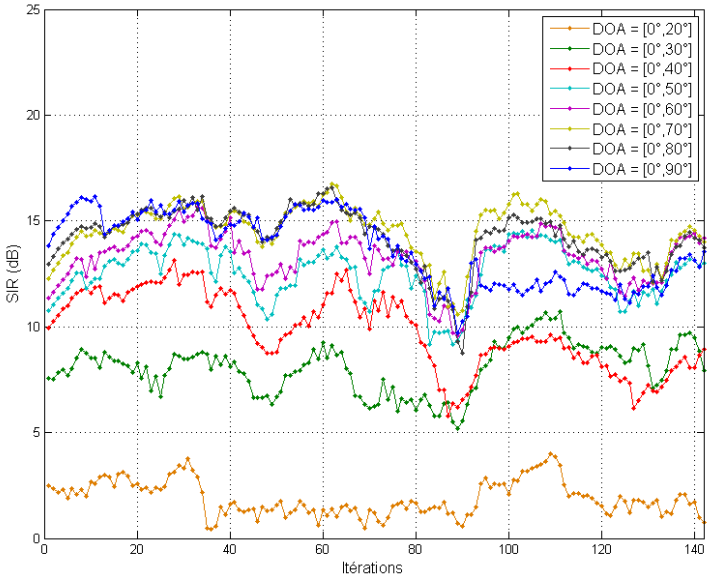


Figure 10.16 – Le rapport source-à-interférences SIR de BF_fixed[5°]_NbSrEstim+BSS-l_1 au cours des fenêtres d'analyse : le nombre de sources est estimé (nombre de sources réel égal à 2).

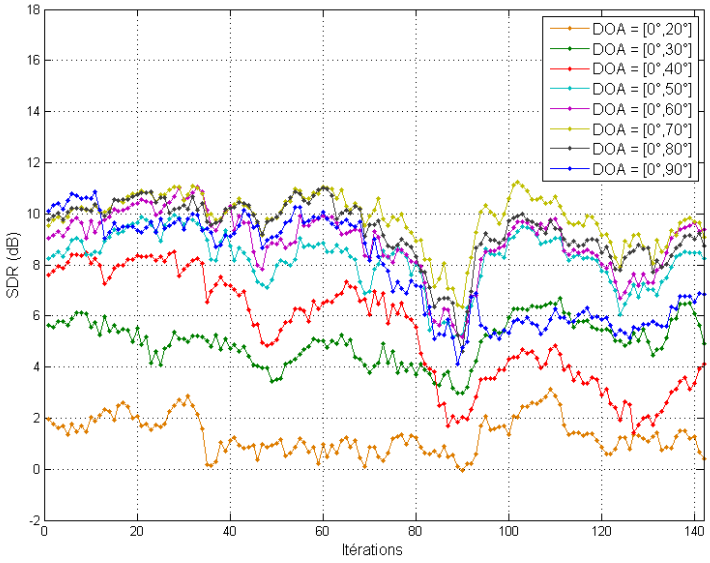


Figure 10.17 – Le rapport source-à-distorsion SDR de BF_fixed[5°]_NbSrEstim+BSS-l_1 au cours des fenêtres d'analyse : le nombre de sources est estimé (nombre de sources réel égal à 2).

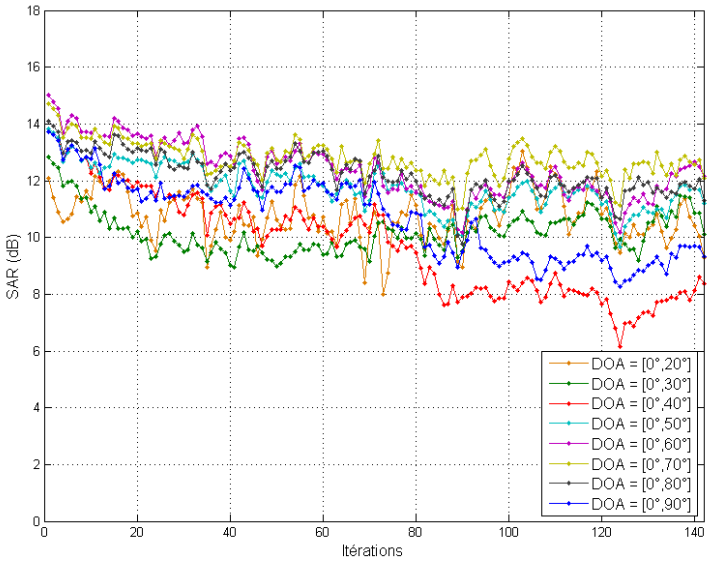


FIGURE 10.18 – Le rapport sources-à-artéfacts SAR de BF_fixed[5°]_NbSrEstim+BSS-l_1 au cours des fenêtres d'analyse : le nombre de sources est estimé (nombre de sources réel égal à 2).

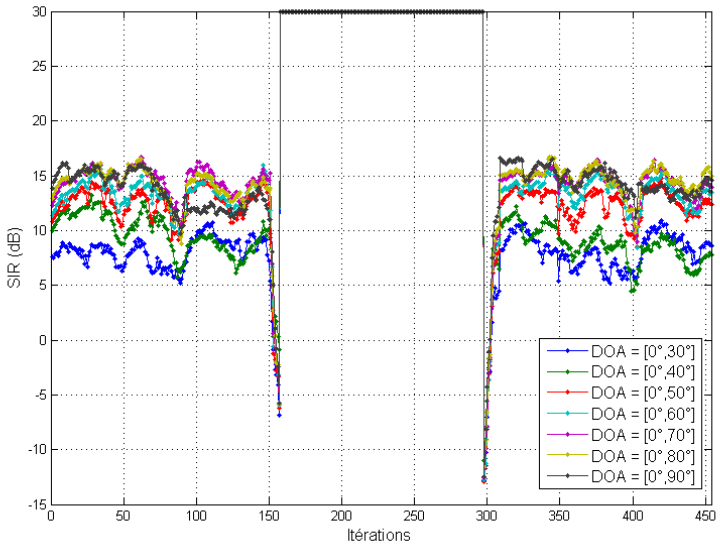


FIGURE 10.19 – Le rapport source-à-interférences SIR de BF_fixed[5°]_NbSrEstim+BSS-l_1 au cours des fenêtres d'analyse : le nombre de sources est estimé et variable entre 1 et 2 sources.

Les figures 10.19, 10.20 et 10.21 montrent le SIR, le SDR et le SAR moyens dans les cas de séparation où le nombre de sources est variable au cours du temps et pour

différentes directions d'arrivées.

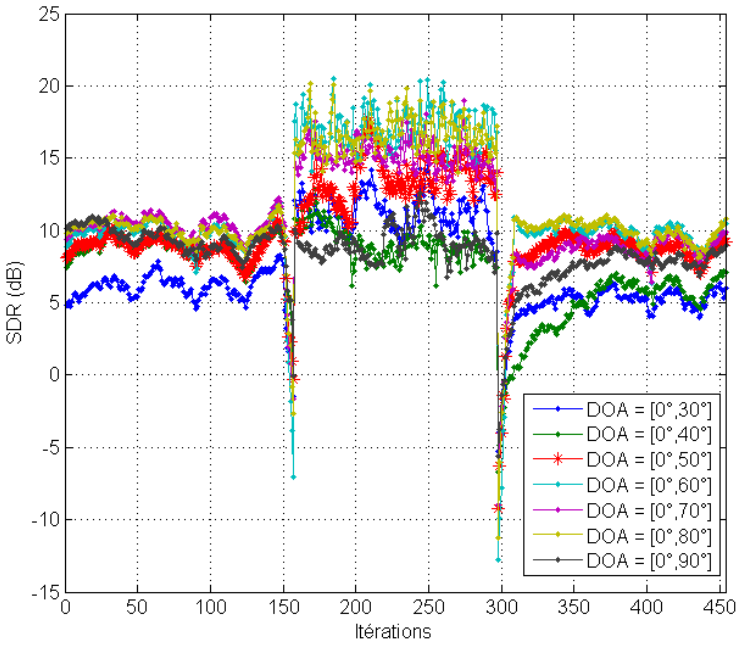


Figure 10.20 – Le rapport source-à-distorsion SDR de BF_fixed[5°]_NbSrEstim+BSS-l_1 au cours des fenêtres d'analyse : le nombre de sources est estimé et variable entre 1 et 2 sources.

Analysons le SIR des sources estimées dans la figure 10.19. Entre la première et la $150^{ème}$ itérations (le terme itération fait référence à la fenêtre d'analyse longue glissante), le SIR estimé pour les directions d'arrivées considérées varie entre 8 et 16 dB et converge correctement. Ensuite, entre les itérations 150 et 157, nous observons une phase de transition se traduisant par une baisse du SIR due à la disparition progressive de la source qui va s'éteindre. Entre les itérations 158 et 299, une seule source est active ce qui se traduit par un fort rapport source-à-interférences. Quand la source qui a disparue pendant les 141 dernières itérations s'active de nouveau, nous assistons à une nouvelle phase de transition qui se traduit par une baisse des performances de séparation. Cette baisse de SIR est due au nombre d'échantillons encore peu nombreux de la source qui vient de s'activer. Le SIR augmente au fur et à mesure que plus d'échantillons de la source qui vient de s'activer sont disponibles, ce qui est la conséquence d'une meilleure séparation. Cette augmentation du SIR se stabilise après la 350 itération pour atteindre la même performance qu'entre la première et la $150^{ème}$ itérations. Notre algorithme suit donc bien le changement dynamique du nombre de sources et converge assez rapidement. Nous rappelons

que la matrice de séparation est initialisée une seule fois et que l'adaptation est totalement automatique et dépend du nombre de sources estimé.

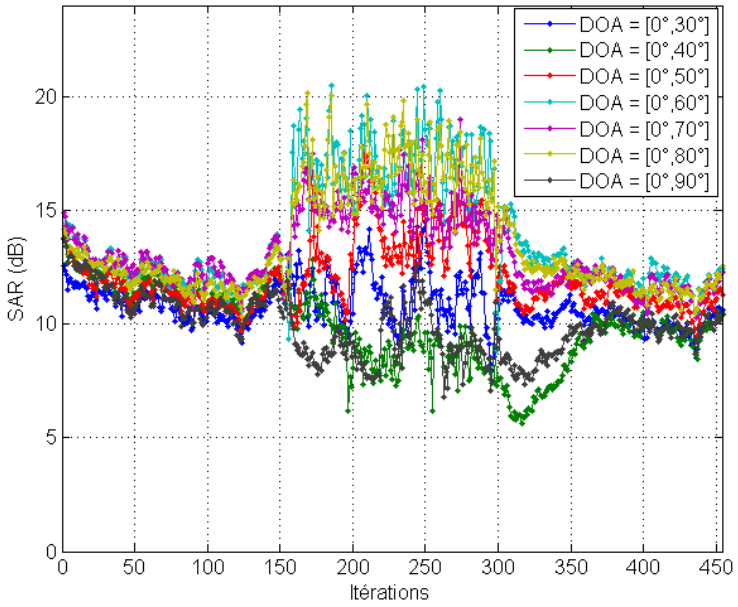


FIGURE 10.21 – Le rapport sources-à-artéfacts SAR de BF_fixed[5°]_NbSrEstim+BSS-l_1 au cours des fenêtres d'analyse : le nombre de sources est estimé et variable entre 1 et 2 sources.

10.3 Résumé des résultats

Pour résumer les résultats obtenus avec notre algorithme de séparation adaptative de sources audio, nous avons calculé la moyenne sur toutes les fenêtres d'analyse et pour chaque direction d'arrivées du rapport source-à-interférences SIR, du rapport source-à-distorsion SDR et du rapport sources-à-artéfacts SAR quand le nombre de sources est fixe (*cf.* figures 10.22, 10.23 et 10.24) et variable (*cf.* figures 10.25, 10.26 et 10.27) au cours du temps. Dans chaque figure, nous avons tracé les performances moyennes des trois cas que nous avons étudiés : le nombre de sources à séparer est connu (courbes bleues), le nombre de sources à séparer est fixé *a priori* ($N_{hyp} = 5$, courbes rouges) et le nombre de sources à séparer est estimé (courbes vertes).

Dans le cas où le nombre de sources est fixe au cours des itérations, les performances de séparation quand le nombre de sources est connu sont les meilleures (*cf.* figures 10.22, 10.23 et 10.24). Cependant les performances de l'algorithme de séparation avec estimation du nombre de sources ne sont pas loin derrière et nous

remarquons qu'il y a entre 1 et 2 dB d'écart entre ces deux algorithmes. Cet écart est dû à une sur-estimation ou une sous-estimation du nombre de sources.

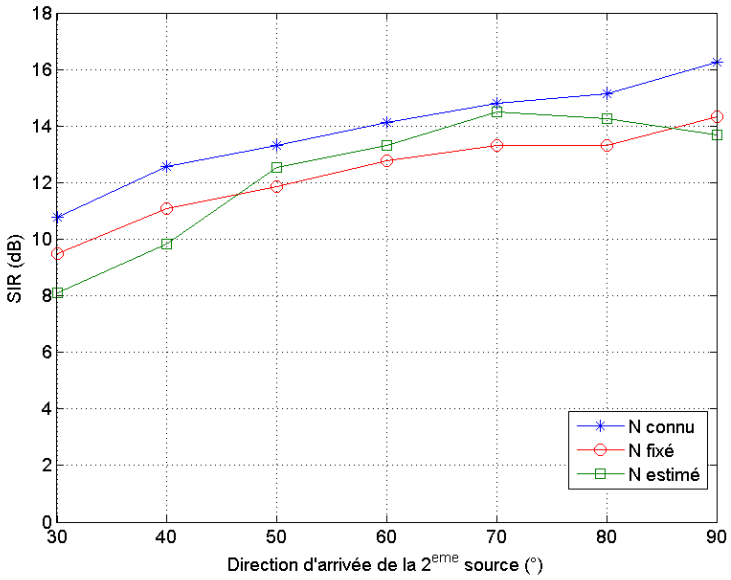


Figure 10.22 – Le rapport source-à-interférences SIR moyen sur toutes les fenêtres d'analyse longues : nombre de sources fixe au cours des itérations.

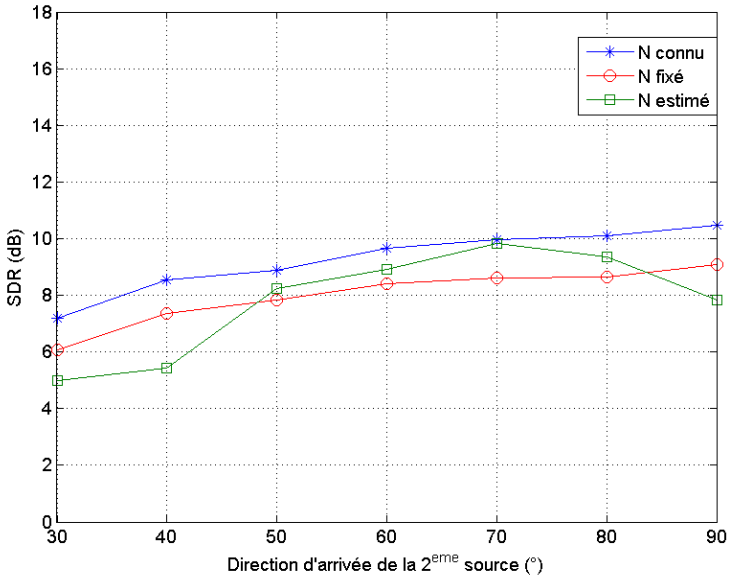


Figure 10.23 – Le rapport source-à-distorsion SDR moyen sur toutes les fenêtres d'analyse longues : nombre de sources fixe au cours des itérations.

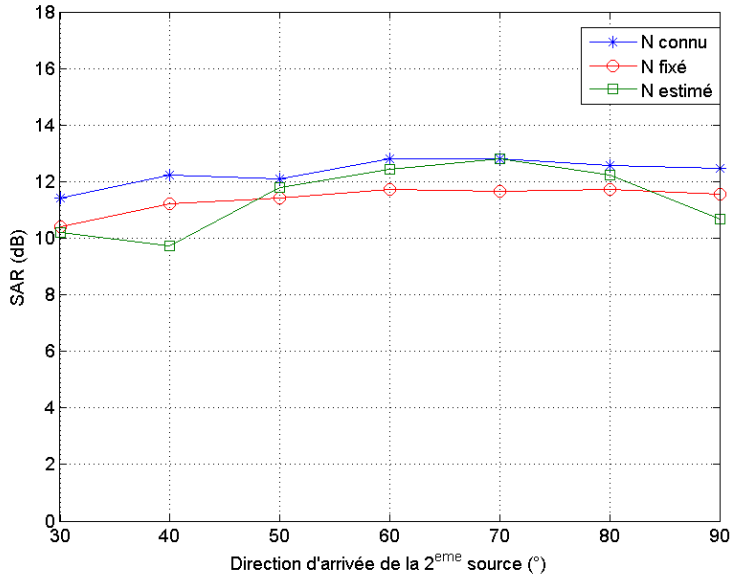


FIGURE 10.24 – Le rapport sources-à-artéfacts SAR moyen sur toutes les fenêtres d'analyse longues : nombre de sources fixe au cours des itérations.

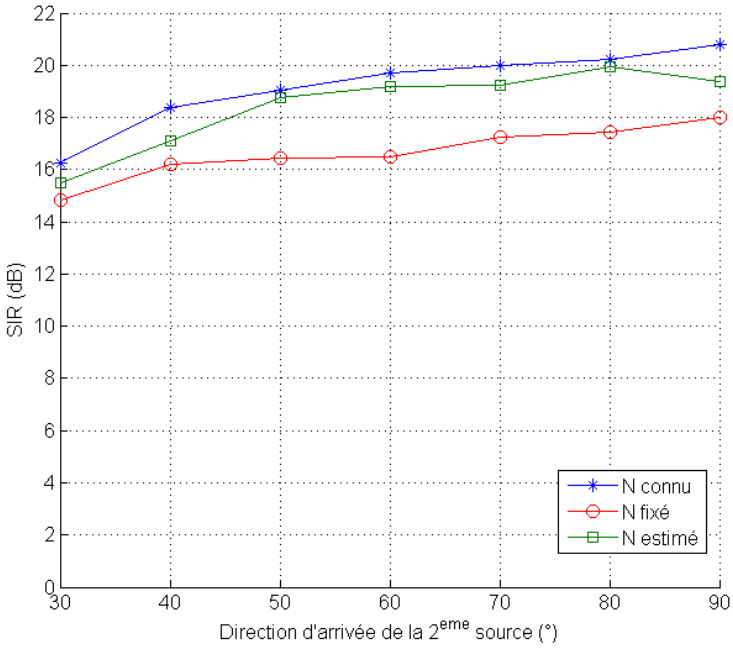


FIGURE 10.25 – Le rapport source-à-interférences SIR moyen sur toutes les fenêtres d'analyse longues : nombre de sources variable au cours des itérations

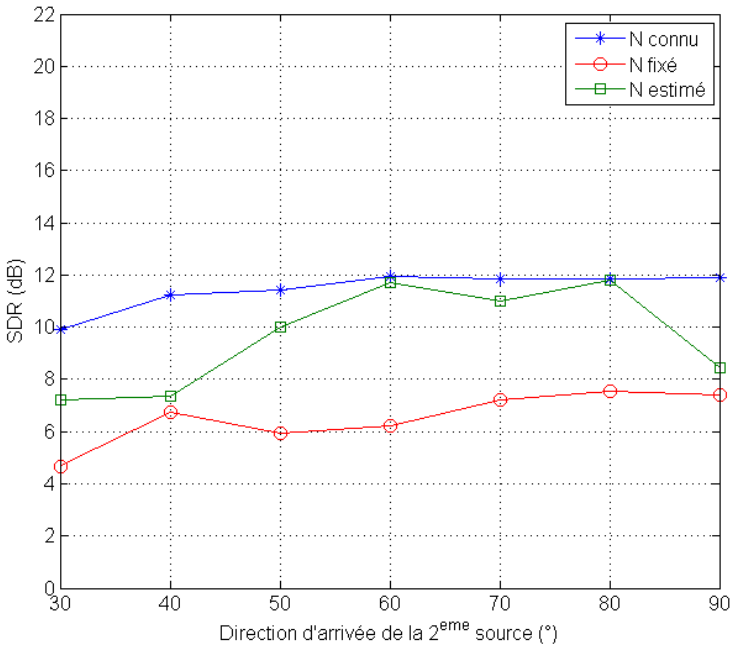


Figure 10.26 – Le rapport source-à-distorsion SDR moyen sur toutes les fenêtres d'analyse longues : nombre de sources variable au cours des itérations

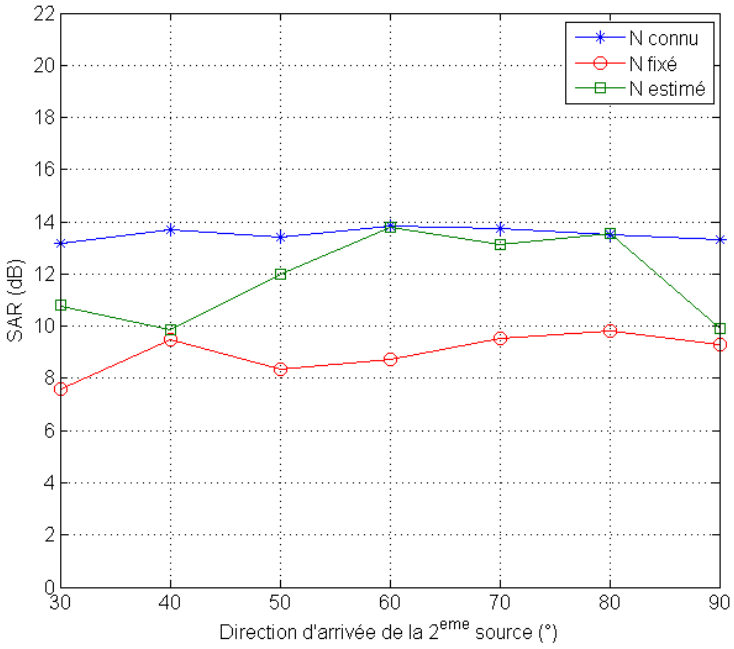


Figure 10.27 – Le rapport sources-à-artéfacts SAR moyen sur toutes les fenêtres d'analyse longues : nombre de sources variable au cours des itérations

10.4 Comparaison avec HARK

10.4.1 HARK : principe

Dans le chapitre consacré à l'état de l'art de la séparation de sources audio, nous avons cité le logiciel d'audition des robots HARK [55]. HARK, pour HRI-JP Audition for Robots with Kyoto University, qui veut dire « listen » en anglais médiéval, est un logiciel open source d'audition des robots. HARK a été réalisé en 2008 et la dernière version 1.0.0 a été introduite en novembre 2010, il se base sur l'environnement de programmation par flot de données Flowdesigner, fonctionne en temps-réel, comporte un certain nombre de modules pour l'audition des robots et supporte les convertisseurs analogiques/numériques multicanaux. Ce logiciel complet d'audition des robots comporte les modules suivants :

1. localisation de sources avec MUSIC (MUltiple SIgnal Classification) ;
2. séparation de sources avec une décorrélation géométrique d'ordre supérieur et un pas d'adaptation adaptatif, cet algorithme de séparation s'appelle GHDSS-AS (Geometric High-order Decorrelation-based Source Separation with Adaptive Stepsize) ;
3. amélioration de la qualité de la parole (*Speech enhancement*) avec HRLE (Histogram-based Recursive Level Estimation) ;
4. reconnaissance automatique de la parole avec MFT-ASR.

Nous nous intéressons au module de séparation de sources GHDSS-AS afin de le comparer à notre algorithme de séparation adaptatif avec estimation du nombre de sources BF_fixed[5°]_NbSrEstim+BSS-l_1.

GHDSS-AS [61] est un algorithme de séparation de sources hybride qui combine formation de voies et séparation de sources, mais d'une manière complètement différente de la nôtre, et qui utilise un pas d'adaptation adaptatif.

Soient un signal audio de N sources $\mathbf{s}(t) = [s_1(t), \ldots, s_N(t)]^T$ et d'un réseau de M microphones. Les sorties du réseau de capteurs sont notés $\mathbf{x}(t) = [x_1(t), \ldots, x_M(t)]^T$, t étant l'indice de temps. Soit $\mathbf{h} = [\mathbf{h}(0), \ldots, \mathbf{h}(L-1)]$ le vecteur des réponses impulsionnelles tronquées à la longueur L, avec $\mathbf{h}(l) = [h_{ij}]_{\ 1 \leq i \leq N,\, 1 \leq j \leq M}$ une matrice de dimension $N \times M$ contenant les $l^{ième}$ coefficients des réponses impulsionnelles des différents chemins acoustiques entre les N sources et M capteurs. Les mélanges à la sortie des capteurs sont la somme des convolutions entre les signaux sources et les réponses impulsionnelles

des différents chemins de propagation entre les sources et les capteurs. Dans le domaine temps-fréquence, les signaux du mélange à l'indice fréquentiel f et à la trame temporelle k peuvent être approximés par :

$$\mathbf{X}\left(f,k\right) \simeq \mathbf{H}\left(f\right)\mathbf{S}\left(f,k\right) \tag{10.1}$$

$\mathbf{X}(f,k) = [X_1(f,k),\ldots,X_M(f,k)]^H$ (respectivement $\mathbf{S}(f,k) = [S_1(f,k),\ldots,S_N(f,k)]^H$) est la TFCT de$\{\mathbf{x}\left(t\right)\}_{1\leq t\leq T}$ (respectivement de $\{\mathbf{s}\left(t\right)\}_{1\leq t\leq T}$) à la fréquence $f \in \left[1, \frac{N_f}{2}+1\right]$ et la fenêtre d'analyse $k \in [1, N_T]$. $\mathbf{H}$ est la TFCT des filtres de mélanges $\{\mathbf{h}\left(l\right)\}_{0\leq l\leq L-1}$. Le but est toujours de trouver, à chaque fréquence, une matrice de séparation $\mathbf{W}\left(f\right)$ qui conduira à l'estimation des sources originales dans le domaine temps-fréquence :

$$\mathbf{Y}\left(f,k\right) = \mathbf{W}\left(f\right)\mathbf{X}\left(f,k\right) \tag{10.2}$$

avec $\mathbf{Y}(f,k) = [Y_1(f,k),\ldots,Y_N(f,k)]^H$. Cette matrice de séparation est estimée en minimisant une fonction de coût $J(\mathbf{W})$ par une méthode de descente de gradient par exemple, l'équation de mise à jour s'écrit :

$$\mathbf{W}_{n+1} = \mathbf{W}_n - \mu\nabla J\left(\mathbf{W}_n\right) \tag{10.3}$$

où $\mathbf{W}_n$ représente $\mathbf{W}$ à l'itération n et μ est le pas d'avancement.

La méthode de séparation de sources proposée par Nakajima et *al.* [61] consiste en la combinaison d'un algorithme de séparation de sources basé sur la décorrélation d'ordre supérieur (HDSS : High-Order Decorrelation based Source Separation) et d'une contrainte géométrique typiquement par formations de voies. La fonction de coût globale J_{GHDSS} s'écrit alors somme suit :

$$J_{GHDSS} = \alpha J_{HDSS}(\mathbf{W}) + \beta J_{GC}(\mathbf{W}) \tag{10.4}$$

où J_{HDSS} est la fonction de coût relative à la décorrélation d'ordre supérieur GHDSS, J_{GC} est la fonction de coût relative à la contrainte géométrique et α et β sont deux facteurs de poid tel que $\alpha+\beta=1$.

HDSS est un algorithme qui utilise une matrice de corrélation d'ordre supérieur $\mathbf{E}_\phi = \phi(\mathbf{Y}\left(f,:\right))\mathbf{Y}^H\left(f,:\right) - \mathrm{diag}\left(\phi(\mathbf{Y}\left(f,:\right))\mathbf{Y}^H\left(f,:\right)\right)$ comme fonction de coût, où ϕ est une fonction non linéaire. $J\left(\mathbf{W}_n\right)$ pour HDSS est définit comme suit :

$$J_{HDSS}\left(\mathbf{W}\right) = \left\|E\left(\mathbf{E}_\phi\right)\right\|^2 \tag{10.5}$$

La contrainte géométrique est basée sur une formation de voies de type *Delay-and-Sum* et s'écrit comme suit :

$$J_{GC}\left(\mathbf{W}_n\right) = \left\|E\left(\mathbf{E}_{GC}\right)\right\|^2 \tag{10.6}$$

$$\mathbf{E}_{GC} = \mathrm{diag}\left(\mathbf{WD} - \mathbf{I}\right) \tag{10.7}$$

Dans [61], les auteurs proposent aussi d'utiliser des pas de mise à jour μ_{HDSS} et μ_{GC} relatifs à chcune des deux fonctions de coût 10.5 et 10.6. Ces pas de mise à jour sont calculés d'une manière adaptative de la manière suivante :

$$\mu_{HDSS} = \frac{\left\|\mathbf{E}_{\phi}\right\|^2}{2\left\|2\mathbf{E}_{\phi}\tilde{\phi}\left(\mathbf{Y}\left(f,:\right)\right)\mathbf{X}^H\left(f,:\right)\right\|^2} \tag{10.8}$$

$$\mu_{GC} = \frac{\left\|\mathbf{E}_{GC}\right\|^2}{2\left\|\mathbf{E}_{GC}\mathbf{D}^H\right\|^2} \tag{10.9}$$

où $\tilde{\phi}(Y_i(f,k)) = \phi\left(Y_i(f,k)\right) + Y_i(f,k)\frac{\partial\phi(Y_i(f,k))}{\partial Y_i(f,k)}$.

10.4.2 Evaluation des résultats

Nous avons comparé les performances de séparation de notre algorithme adaptatif de séparation de source avec prétraitement par formation de voies fixe suivi de la sélection de lobes et estimation du nombre de sources BF_fixed[5°]_NbSrEstim+BSS-l_1 avec celle obtenues par le module de séparation de source GHDSS-AS de HARK (nommé HARK dans les figures 10.28 et 10.29). Cette évaluation des performances de séparation a été effectué sur 40 cas de séparation de deux sources de la base de données Theo-RI-studio, la première source est placée à 0° et la deuxième à 60°, le nombre de source est fixe au cours du temps. Pour une source estimée, les mesures de qualité ont été évaluées sur la totalité de ce signal, donc après sa reconstruction, et non pas sur les fenêtres d'analyse comme nous l'avons effectué dans les deux sections précédentes. Les mesures qui seront présentées dans la suite sont la moyenne et l'écart type des 40 cas de séparation de deux sources.

La figure 10.28 montre que les performances de séparation de notre algorithme BF_fixed[5°]_NbSrEstim+BSS-l_1 en terme de rapport source-à-interférences SIR, de rapport source-à-distorsion SDR et de rapport sources-à-artéfacts SAR sont au même niveau que celles obtenues par HARK.

La figure 10.29 montre les performances de séparation perceptuelles en terme

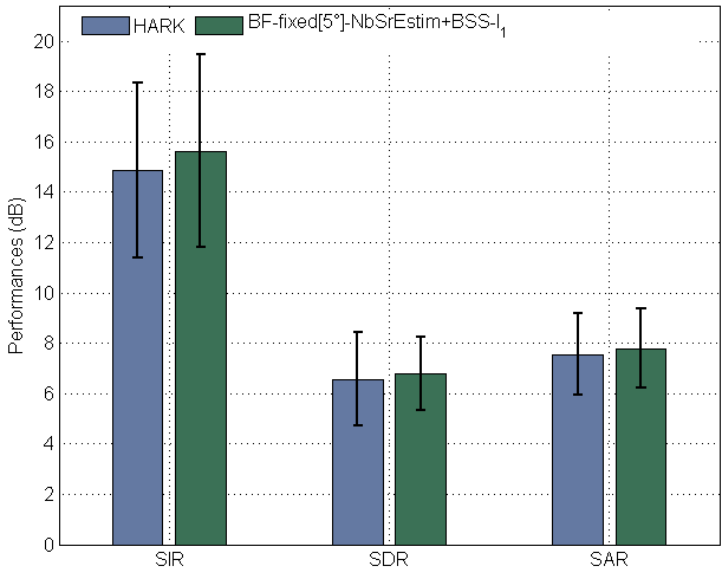


Figure 10.28 – Le rapport source-à-interférences (SIR), le raport source-à-distorsion (SAR) et le rapport sources-à-artéfacts (SAR) des sources estimées avec HARK (en bleu) et BF_fixed[5°]_NbSrEstim+BSS-l_1 (en vert) et leurs écarts types (barres verticales) : les moyennes pour la séparation de 40 paires de sources de la base de données Theo-RI-studio, les directions d'arrivées sont 0° et 60°.

de score perceptuel global OPS, de score perceptuel relatif à la source cible TPS, de score perceptuel relatif aux interférences IPS et de score perceptuel relatif aux artéfacts APS de BF_fixed[5°]_NbSrEstim+BSS-l_1 et de HARK. Si nous analysons ces mesures, nous remarquons que :

- en terme de préservation de la source cible (TPS), nous avons un score proche de celui obtenu par HARK ;
- en terme de suppression des autres sources, celles qui sont différentes de la source cible (IPS), nous avons un score supérieur à celui de HARK d'environ 40 points ;
- en terme d'absence de bruit artificiel additionnel dans les sources estimées (APS), nous avons un score inférieur à celui de HARK de 25 points.

Les observations précédentes nous amènent à un score perceptuel global de notre algorithme BF_fixed[5°]_NbSrEstim+BSS-l_1 inférieur à celui de HARK de 15 points.

Ces premiers résultats obtenus en comparant notre algorithme de séparation aveugle de sources avec estimation du nombre de sources BF_fixed[5°]_NbSrEstim+BSS-l_1 à l'algorithme référence de l'état de l'art relatif à la séparation de sources pour l'audition des robots sont très prometteurs. Ils montrent que les résultats en terme

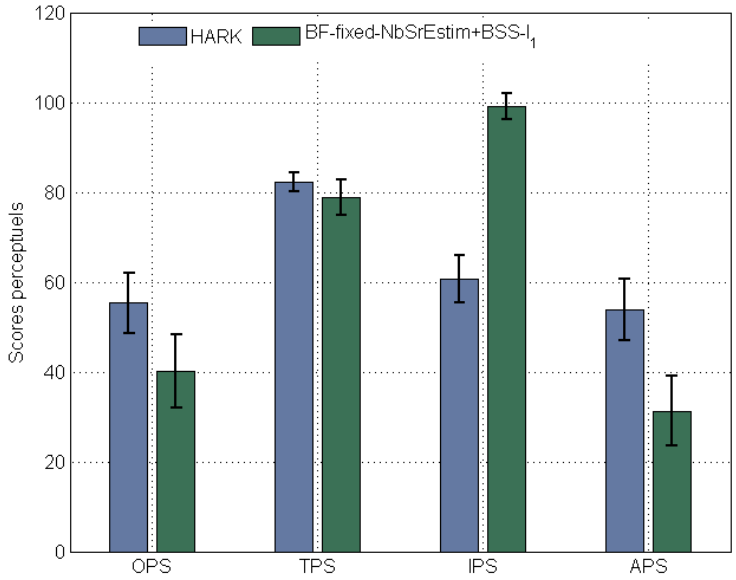


FIGURE 10.29 – Le score perceptuel global (OPS), le score perceptuel relatif à la cible (TPS), le score perceptuel relatif aux interférences (IPS) et le score perceptuel relatif aux artéfacts (APS) des sources estimées avec HARK (en bleu) et BF_fixed[5°]_NbSrEstim+BSS-l_1 (en vert) et leurs écarts types (barres verticales) : les moyennes pour la séparation de 40 paires de sources de la base de données Theo-RI-studio, les directions d'arrivées sont 0° et 60°.

de séparation de sources sont comparables à ceux obtenus par HARK et qu'un travail reste à faire afin d'améliorer le score perceptuel relatif aux artéfacts.

Conclusion

Dans ce chapitre, nous nous sommes intéressés à la séparation de sources adaptative avec un nombre de sources fixe ou variable au cours du temps. Nous avons testé trois configurations : le nombre de sources est connu, le nombre de sources est inconnu mais fixé *a priori* à une valeur maximale choisie (*i.e.* sur-estimé) et le nombre de sources est estimé. Les conclusions suivantes peuvent être tirées :

- les performances de l'algorithme de séparation avec estimation du nombre de sources BF_fixed[5°]_NbSrEstim+BSS-l_1 sont proches de celles où le vrai nombre de sources est connu BF_fixed[5°]_BS+BSS-l_1 de 2dB en moyenne ;
- Dans le cas où nous estimons un nombre de sources supérieur au nombre de sources réel, les performances de séparation sont affectées par des distorsions et des artéfacts ;

– notre algorithme BF_fixed[5°]_NbSrEstim+BSS-l_1 a des performances très proches de celles de l'algorithme référence de l'état de l'art de l'audition des robots HARK, les performances de séparation ont été évaluées pour la séparation de 40 paires de sources de la base de données Theo-RI-studio, les directions d'arrivées sont 0° et 60°.

Chapitre 11

Conclusion et perspectives

Réalisations

Dans cette thèse, nous avons proposé de nouveaux algorithmes de séparation aveugle de sources audio avec un réseau de microphones pour l'application à l'audition des robots.

D'abord, nous avons proposé un algorithme de séparation de sources basé sur un critère de parcimonie qui est la minimisation de la norme l_1 de la représentation temps-fréquence des signaux séparés ; cet algorithme utilise comme méthode d'optimisation l'algorithme du gradient naturel. La comparaison avec l'analyse en composantes indépendantes qui minimise l'information mutuelle en utilisant la même méthode d'optimisation montre que la minimisation de la norme l_1 a les même performances moyenne qu'ICA. Ceci nous a mené à étudier l'effet de la modification de la contrainte de parcimonie sur les performances de la séparation de sources. Nous avons développé un algorithme de séparation de sources basé sur la minimisation de la pseudo-norme l_p, $0 < p < 1$. Nous avons rendu la contrainte de parcimonie de plus en plus dure au fur et à mesure que l'algorithme avance dans ses itérations et ceci en faisant décroître le paramètre p de 1 à une valeur proche de 0 selon une fonction sigmoïde. Cette méthode de séparation donne de meilleurs résultats que la minimisation de la norme l_1.

Ensuite, nous avons proposé une classe d'algorithmes de séparation à deux étapes : une étape de prétraitement par formation de voies fixe et une étape de séparation de sources. Nous avons présenté et testé différentes configurations du prétraitement par formation de voies : une formation de voies vers les directions d'arrivées, une formation de voies avec un nombre de lobes fixe et une formation de voies avec un nombre de lobes fixe suivi d'une sélection des lobes contenant les plus grandes

énergies des sources. Pour construire les filtres de la formation de voies en tenant compte de l'effet de la tête du robot sur le champ sonore proche, nous avons utilisé les fonctions de transfert de tête (HRTF) comme vecteurs directionnels. Les résultats montrent que le prétraitement par formation de voies et la sélection des lobes ayant la plus grande énergie reçue améliore de minimum 10dB les résultats de la séparation de sources en termes de SIR, SDR et SAR mais aussi perceptuellement d'environ 40 points.

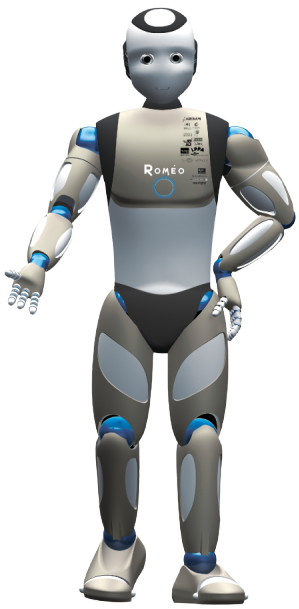
Une partie de cette thèse s'est déroulé dans le laboratoire d'acoustique et le studio d'enregistrement de Télécom ParisTech. Nous avons enregistré une base de données de HRTF pour le calcul des filtres de formation de voies et une base de données de réponses impulsionnelles pour l'évaluation des performances des algorithmes de séparation de sources. Nous avons aussi fait des enregistrements de réponses impulsionnelles dans l'appartement témoin du projet Romeo à l'Institut de la Vision.

Nous avons développé les algorithmes proposés en version itérative et en version adaptative. Nous avons proposé une version adadptative capable de suivre le changement du nombre de sources au cours du temps. En effet, notre algorithme adadptatif BF_fixed[5°]_NbSrEstim+BSS-l_1 est capable d'estimer le nombre de sources actives à chaque fenêtre d'analyse et de mettre à jour la matrice de séparation d'une manière automatique, cette matrice étant initialisée une seule fois au début du traitement. La comparaison de notre algorithme adaptatif avec le module de séparation de sources de l'algorithme de l'audition des robots HARK montre que nous avons des performances très proches, ce qui constitue un résultat prometteur pour la suite.

Nous avons évalué les algorithmes proposés en utilisant deux méthodes d'évaluation, l'une basée sur des critères de performance objectifs (BSS-EVAL) et la deuxième basée sur des critères de performance perceptuels (PEASS).

Tests avec Romeo

Les algorithmes de séparation de sources développés doivent être testés avec l'humanoïde *Romeo* (*cf.* figure 11.1a). Nous avons reçu le prototype de la tête et torse de Romeo (*cf.* figure 11.1b) au mois de novembre 2011 à la fin de ma dernière année de thèse, malheureusement ce prototype souffre encore de problèmes techniques qui nous ont empêchés de faire des mesures et des expériences. Romeo dispose de 16 capteurs dont deux capteurs placés à l'intérieur de deux tubes commençant par deux pavillons qui modélisent les pavillons et les conduits auditifs humains. Les positions



(a) Romeo, l'humanoïde final.

(b) Le prototype tête et torse de Romeo.

FIGURE 11.1 – Romeo

des microphones sur la tête de Romeo sont décrites dans la figure 11.2. Dans un futur très proche, nous enregistrerons avec Romeo la base de données des HRTF et des réponses impulsionnelles afin de tester les algorithmes de séparation de sources que nous avons développés.

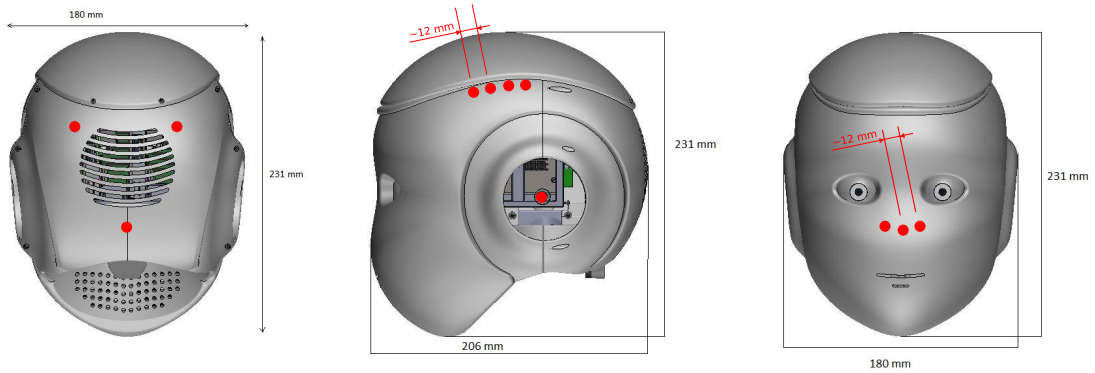


FIGURE 11.2 – La position des capteurs dans la tête de Romeo.

Perspectives

Vers une interaction Homme-Robot multimodale Dans cette thèse, nous avons travaillé sur la séparation de sources pour l'audition des robots d'un point de vue purement traitement de signal audio. De plus en plus de travaux incluent la « vision » pour aider à la séparation de sources : le robot peut localiser les sources à séparer par la ou les caméras dont il dispose. Nous nous dirigeons donc vers une séparation de sources et une interaction Homme-robot multimodale

Vers un apprentissage du milieu sonore Un robot humanoïde destiné à évoluer dans le même environnement (par exemple un appartement) peut apprendre les caractéristiques acoustique et sonore de cet environnement. Cet apprentissage peut être à plusieurs niveaux :

- apprentissage de la localisation des sources interférentes ou les sources de bruits fixes : la télévision, la radio, le bruit d'un ventilateur ou le bruit provenant d'une fenêtre ouverte sont des sources de bruits ou des sources interférentes qui ne bouge pas ou très peu au cours du temps, un apprentissage de leur position permettrait au robot « d'éviter » ces zones où l'acquisition et la séparation des sources utiles peuvent être compromises ;
- apprentissage des sources sonores usuelles (répétitives) : la sonnette de la porte, la sonnerie du téléphone ou bien l'alarme du micro-onde sont des sources sonores utiles, répétitives et ayant des caractéristiques spectrales bien particulières, un apprentissage de ces sources peut faciliter leur séparation de la parole par exemple ;
- apprentissage de l'acoustique des pièces : l'humanoïde pourrait varier la complexité de l'algorithme de séparation de sources en fonction du taux de réverbération de la pièce dans laquelle il se trouve.

Bibliographie

Publications de l'auteur

[1] Slim Essid, Yves Grenier, Mounira Maazaoui, Gael Richard et Robin Tournemenne : An audio-driven virtual dance-teaching assistant. *ACM Multimedia*, 2011.

[2] Mounira Maazaoui, Karim Abed-Meraim et Yves Grenier : Adaptive blind source separation with hrtfs beamforming preprocessing and varying number of sources. *The Sventh IEEE Sensor Array and Multichannel Signal Processing Workshop, SAM 2012*, 2012.

[3] Mounira Maazaoui, Karim Abed-Meraim et Yves Grenier : Blind source separation for robot audition using fixed hrtf beamforming. *EURASIP Journal on Advances in Signal Processing*, 2012.

[4] Mounira Maazaoui, Yves Grenier et Karim Abed-Meraim : Blind source separation for robot audition using fixed beamforming with hrtfs. *21th Annual Conference on the International Speech Communication Association, Interspeech 2011*, 2011.

[5] Mounira Maazaoui, Yves Grenier et Karim Abed-Meraim : Frequency domain blind source separation for robot audition using a parameterized sparsity criterion. *19th European Signal Processing Conference, EUSIPCO 2011*, 2011.

[6] Mounira Maazaoui, Yves Grenier et Karim Abed-Meraim : From binaural to multimicrophone blind source separation using fixed beamforming with hrtfs. *The 19th International Conference on Systems, Signals and Image Processing, IWSSIP 2012*, 2012.

Références

[7] Romeo project : www.projetromeo.com.

[8] R. A. BROOKS et L. ANDREA STEIN : Building brains for bodies. *Autonomous Robots*, pages 7–25, 1994.

[9] B. A. PEARLMUTTER et L. PARRA : Maximum likelihood blind source separation : A context-sensitive generalization of ICA. *Advances in Neural Information Processing Systems*, pages 613–619, 1997.

[10] R. AICHNER, H. BUCHNER, S. ARAKI et S. MAKINO : On-line time-domain blind source separation of nonstationary convolved signals. *International Symposium on Independent component analysis and Blind Signal Separation (ICA)*, 2003.

[11] R. AICHNER, H. BUCHNER, F. YAN et W. KELLERMANN : A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments. *Signal Processing*, pages 1260 – 1277, 2006.

[12] V.R. ALGAZI, R.O. DUDA, D.M. THOMPSON et C. AVENDANO : The CIPIC HRTF database. *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 99 –102, 2001.

[13] S. AMARI, A. CICHOCKI et H. H. YANG : A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems*, pages 757–763, 1996.

[14] S.-I. AMARI : Natural gradient works efficiently in learning. *Neural Computation*, pages 251–276, 1998.

[15] J. BENESTY, J. CHEN et Y. HUANG : *Microphone Array Signal Processing, Chapter 3 : Conventional beamforming techniques*. Springer, 1rst édition, 2008.

[16] J. BLAUERT : *Spatial hearing, the psychophysics of human sound localization*. MIT Press, 1983.

[17] A. S. BREGMAN : *Auditory scene analysis*. MIT Press, Cambridge MA, 1990.

[18] H. BUCHNER, R. AICHNER et W. KELLERMANN : A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. *IEEE Transaction on Speech audio processing*, pages 120–134, 2005.

[19] V. Capdevielle, C. Serviere et J.L. Lacoume : Blind separation of wide-band sources in the frequency domain. *International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1995*, pages 2080–2083, 1995.

[20] J.-F. Cardoso : Blind signal separation : statistical principles. *Proceedings of the IEEE*, pages 2009–2025, 1998.

[21] J.F. Cardoso et A. Souloumiac : Blind beamforming for non-Gaussian signals. *Radar and Signal Processing*, pages 362 –370, 1993.

[22] P. Comon : Independent component analysis. *Higher Order Statistics, J.L. Lacoume (Editor), Elsevier*, pages 29–38, 1992.

[23] P. Comon : Independent component analysis, a new concept ? *Signal Processing*, 1994.

[24] P. Comon et C. Jutten : *Handbook of Blind Source Separation , Independent Component Analysis and Applications*. Elsevier, 2010.

[25] P. Comon et L. Rota : Blind separation of independent sources from convolutive mixtures. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, pages 542–549, 2003.

[26] P. D. O grady, B. A. Pearlmutter et S. T. Rickard : Survey of sparse and non-sparse methods in source separation. *International Journal of Imaging Systems and Technology*, pages 18–33, 2005.

[27] Ircam AKG HRTF database : http ://recherche.ircam.fr/equipes/salles/listen/index.html, 2003.

[28] Shimada Laboratory HRTF Database : http ://audio.nagaokaut.ac.jp/hrtf/, 2005. URL `http://audio.nagaokaut.ac.jp/hrtf/index_e.html`.

[29] S.C. Douglas et M. Gupta : Scaled natural gradient algorithms for instantaneous and convolutive blind source separation. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 637–640, 2007.

[30] V. Emiya, E. Vincent, N. Harlander et V. Hohmann : Subjective and objective quality assessment of audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, pages 2046 –2057, 2011.

[31] R. F. LYON : A computational model of binaural localization and separation. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1148–1151, 1983.

[32] FFADO : http ://www.ffado.org/ ?q=release/beta.

[33] S. FOSTER : Impulse response measurement using Golay codes. *In IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '86*, pages 929 – 932, 1986.

[34] M. G. PARKER, K. G. PATERSON et C. TELLAMBURA : Golay complementary sequences. *Encyclopedia of telecommunications*, 2004.

[35] B. GARDNER et K. MARTIN : HRTF measurements of a KEMAR dummy-head microphone. Rapport technique, MIT Media Lab Perceptual Computing, 1994.

[36] Yves GRENIER : *Cours Acoustique des salles de l'unité d'enseignement d'acoustique (SI220) à Télécom ParisTech.*

[37] R.R. GUDDETI et B. MULGREW : Perceptually motivated blind source separation of convolutive mixtures. *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2005*, pages 273–276, 2005.

[38] Y.-X. HE, G. XU et Q.-R. QIU : A new model for robot audition using independent component analysis and time-frequency representation. *International Conference on Machine Learning and Cybernetics*, pages 705–709, 2007.

[39] J. HUANG, N. OHNISHI et N. SUGIE : Building ears for robots : sound localization and separation. *Artificial Life and Robotics*, pages 157–163, 1997.

[40] N. HURLEY et S. RICKARD : Comparing measures of sparsity. *IEEE Workshop on Machine Learning for Signal Processing*, pages 4723–4741, 2009.

[41] S. ICART et R. GAUTIER : Blind separation of convolutive mixtures using second and fourth order moments. *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1996*, pages 3018–3021, 1996.

[42] A. J. BELL et T. J. SEJNOWSKI : An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, pages 1129–1159, 1995.

[43] C. Jutten et J. Herault : Blind separation of sources, part 1 : an adaptive algorithm based on neuromimetic architecture. *Signal Processing*, pages 1–10, 1991.

[44] M. Kawamoto, K. Barros, K. Matsuka et N. Ohnishi : A method of real-world blind separation implemented in frequency domain. *Independent Component Analysis 2000*, pages 267–272, 2000.

[45] M. Kawamoto Allan, A. Kardec Barros, A. Mansour, K. Matsuoka et N. Ohnishi : Blind separation for convolutive mixtures of non-stationary signals. *International Conference Neural Inf. Proc*, pages 743–746, 1998.

[46] JACK Audio Connection Kit : http ://qjackctl.sourceforge.net/.

[47] R. Huber & B. Kollmeier : Pemo-q – a new method for pbjective audio quality assessment using a model of auditory perception. *IEEE transactions on audio, speech and language processing*, pages 1902–1911, 2006.

[48] J. Luo et Z. Zhang : Using eigenvalue grads methods to estimate the number of source. *International Conference on Software Process, ICSP*, 2000.

[49] S. Makino, S. Araki, R. Mukai et H. Sawada : Audio source separation based on independent component analysis. *The International Symposium on Circuits and Systems*, pages 668–671, 2004.

[50] S. Makino, T.-W. Lee et H. Sawada : *Blind Speech Separation.* Springer, 2007.

[51] K. Matsuoka, M. Ohya et M. Kawamoto : A neural net for blind separation of nonstationary signals. *Neural Network*, pages 411–419, 1995.

[52] Y. Matsusaka, T. Tojo, S. Kubota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano et T. Kobayashi : Multi-person conversation via multi-model interface - A robot who communicate with multi-user. *6th European Conference on Speech Communication and Technology, EUROSPEECH*, 1999.

[53] L. Molgedey et H. G. Schuster : Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, pages 3634–3637, 1994.

[54] K. NAKADAI, H. G. OKUNO et H. KITANO : Real-time sound source localization and separation for robot audition. *IEEE International Conference on Spoken Language Processing*, pages 193–196, 2002.

[55] K. NAKADAI, H. G. OKUNO, T. TAKAHASHI, K. NAKAMURA, T. MIZUMOTO, T. YOSHIDA, T. OTSUKA et G. INCE : Introduction to open source robot audition software HARK. *The 29th annual conference of the robotics society of Japan, RSJ 2011*, 2011.

[56] K. NAKADAI, T. LOURENS, H. G. OKUNO et H. KITANO : Active audition for humanoid. *The 17th National Conference on Artificial Intelligence*, pages 832–839, 2000.

[57] K. NAKADAI, D. MATSUURA, H.G. OKUNO et H. KITANO : Applying scattering theory to robot audition system : robust sound source localization and extraction. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1147–1152, 2003.

[58] K. NAKADAI, H. NAKAJIMA, H. YUJI et T. HIROSHI : Sound source separation of moving speakers for robot audition. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3685–3688, 2009.

[59] K. NAKADAI, H.G. OKUNO, H. NAKAJIMA, Y. HASEGAWA et H. TSUJINO : An open source software system for robot audition hark and its evaluation. *International Conference onHumanoid Robots*, pages 561 –566, 2008.

[60] H. NAKAJIMA, K. NAKADAI, Y. HASEGAWA et H. TSUJINO : High performance sound source separation adaptable to environmental changes for robot audition. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2165–2171, 2008.

[61] H. NAKAJIMA, K. NAKADAI, Y. HASEGAWA et H. TSUJINO : Blind source separation with parameter-free adaptive step-size method for robot audition. *IEEE Transaction on Audio, Speech, and Language Processing*, pages 1476–1484, 2010.

[62] H.G. OKUNO, T. OGATA et K. KOMATANI : Robot audition from the viewpoint of computational auditory scene analysis. *International Conference on Informatics Education and Research for Knowledge-Circulating Society*, pages 35–40, 2008.

[63] L. Parra et C.V. Alvino : Geometric source separation : merging convolutive source separation with geometric beamforming. *IEEE Transactions on Speech and Audio Processing*, pages 352–362, 2002.

[64] L. Parra et C. Spence : Convolutive blind separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing*, pages 320 – 327, 2000.

[65] L. Parra et C. Spense : On-line blind source separation of non stationary signals. *IEEE Neural Networks and Signal Processing Workshop*, 2000.

[66] M. S. Pedersen, J. Larsen, U. Kjems et L. Parra : A survey of convolutive blind source separation methods. *Springer Handbook of Speech Processing*, 2007.

[67] M.S. Pedersen, U. Kjems, K.B. Rasmussen et L.K. Hansen : Semi-blind source separation using head-related transfer functions. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 713–716, 2004.

[68] D-T. Pham et J-F. Cardoso : Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. Signal Processing*, pages 1837–1848, 2001.

[69] S. Rickard : Sparse sources are separated sources. *The 16th Annual European Signal Processing Conference*, September 2006.

[70] B. Rudzyn, W. Kadous et C. Sammut : Real time robot audition system incorporating both 3D sound source localisation and voice characterisation. *IEEE International Conference on Robotics and Automation*, pages 4733 – 4738, 2007.

[71] T. Rutkowski et A. Cichocki : Speech enhancement from interfering sounds using casa techniques and blind source separation. pages 728–733, 2001.

[72] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa et K. Shikano : Blind source separation combining independent component analysis and beamforming. *EURASIP Journal on Applied Signal Processing*, pages 1135–1146, 2003.

[73] H. Saruwatari, Y. Mori, T. Takatani, S. Ukai, K. Shikano, T. Hiekata et T. Morita : Two-stage blind source separation based on ICA and binary masking for real-time robot audition system. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2303–2308, 2005.

[74] T. SAWADA, T. SEKIYA et T. KOBOYASHI : Recognition of the mixed speech based on multi-stage audio segregation. *The 18th meeting of special interest group on AI challenges*, 2003.

[75] C. SIMON, P. LOUBATON, C. VIGNAT, C. JUTTEN et G. D'URSO : Separation of a class of convolutive mixtures : a contrast function approach. *Signal Processing*, pages 883–887, 1999.

[76] A. SOULOUMIAC : Blind source detection and separation using second order non-stationarity. *International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2005*, pages 1912–1915, 1995.

[77] X. SUN et S.C. DOUGLAS : Adaptive paraunitary filter banks for contrast-based multichannel blind deconvolution. *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2001*, pages 2753–2756, 2001.

[78] T. TAKAHASHI, K. NAKADAI, K. KOMATANI, T. OGATA et H. G. OKUNO : Improvement in listening capability for humanoid robot HRP-2. *IEEE International Conference on Robotics and Automation, ICRA 2010*, pages 470–475, 2010.

[79] T. TAKATANI, S. UKAI, T. NISHIKAWA, H. SARUWATARI et K. SHIKANO : Blind sound scene decomposition for robot audition using SIMO-Model-Based ICA. *IEEE/RSJ International Conference on Intelligent Robot and Systems*, pages 2247–2252, 2005.

[80] Y. TAMAI, S. KAGAMI, Y. AMEMIYA, Y. SASAKI, H. MIZOGUCHI et T. TAKANO : Circular microphone array for robot's audition. *Proceedings of IEEE Sensors*, pages 565–570, 2004.

[81] Y. TAMAI, Y. SASAKI, S. KAGAMI et H. MIZOGUCHI : Three ring microphone array for 3D sound localization and separation for mobile robot audition. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4172–4177, 2005.

[82] J. THOMAS, Y. DEVILLE et S. HOSSEINI : Time-domain fast fixed-point algorithms for convolutive ICA. *IEEE Signal Processing Letters*, pages 228–231, 2006.

[83] F. Tordini et F. Piazza : A semi-blind approach to the separation of real world speech mixtures. *International Joint Conference on Neural Networks, IJCNN 2002*, pages 1293–1298, 2002.

[84] K. Torkkola : Blind separation of convolved sources based on information maximization. *IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing*, pages 423–432, 1996.

[85] J.-M. Valin, J. Rouat et F. Michaud : Enhanced robot audition based on microphone array soure separation with post-filter. *IEEE/RSJ International conference on intelligent robots and systems*, pages 2123 – 2128, 2004.

[86] J.-M. Valin, S.-I. Yamamoto, J. Rouat, F. Michaud, K. Nakadai et H. G. Okun : Robust recognition of simultaneous speech by a mobile robot. *IEEE Transaction on Robotics*, pages 742–752, 2007.

[87] B.D. Van Veen et K.M. Buckley : Beamforming : a versatile approach to spatial filtering. *IEEE ASSP Magazine*, pages 4–24, 1988.

[88] E. Vincent, R. Gribonval et C. Fevotte : Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, pages 1462 –1469, 2006.

[89] L. Wang, H. Ding et F. Yin : Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010.

[90] W. Wang, J. A. Chambers et S. Sanei : A novel hybrid approach to the permutation problem of frequency domain blind source separation. *International conference on independent component analysis and blind signal separation*, pages 532–539, 2004.

[91] W. Wang, J. A. Chambers et S. Sanei : Penalty function approach for constrained convolutive blind source separation. *International conference on independent component analysis and blind signal separation*, pages 661–668, 2004.

[92] E. Weinstein, M. Feder et A.V. Oppenheim : Multi-channel signal separation by decorrelation. *IEEE Transactions on Speech and Audio Processing*, pages 405–413, 1993.

[93] H. C. WU et J. C. PRINCIPE : Simultaneous diagonalization in the frequency domain (SDIF) for source separation. *Independent component analysis*, pages 245–250, 1999.

[94] S. YAMAMOTO, K. NAKADAI, M. NAKANO, H. TSUJINO, J.-M. VALIN, K. KOMATANI, T. OGATA et H.G. OKUNO : Design and implementation of a robot audition system for automatic speech recognition of simultaneous speech. *IEEE Workshop on Automatic Speech Recognition Understanding*, pages 111–116, 2007.

[95] S.-I. YAMAMOTO, K. NAKADAI, M. NAKANO, H. TSUJINO et J.-M. VALIN : Real-time robot audition system that recognize simultaneous speech in the real world. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5333–5338, 2006.

[96] Zhongfu YE, Chunqi CHANG, Chen WANG, Jian ZHAO et F.H.Y. CHAN : Blind separation of convolutive mixtures based on second order and third order statistics. *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2003*, pages 305–308, 2003.

Liste des algorithmes

4.1 Algorithme de la minimisation de la norme l_1 62
4.2 Algorithme de la minimisation de la pseudo-norme l_p paramétrée . . . 66
5.1 Algorithme général de séparation de sources avec prétraitement de formation de voies 71
5.2 Formation de voies vers les directions d'arrivées BF_DOA 73
5.3 Formation de voies fixe BF_fixed 74
5.4 Formation de voies fixe avec sélection de lobes, le nombre de source est supposé connu *a priori,* BF_fixed_BS 79
5.5 Formation de voies fixe avec sélection de lobes et estimation du nombre de sources, BF_fixed_NbSrEstim 80
6.1 Séparation adaptative de sources avec prétraitement de formation de voies : le nombre de sources est connu *a priori* ou inconnu et fixé arbitrairement, BF_fixed_BS+BSS 87
6.2 Séparation adaptative de sources avec prétraitement de formation de voies et estimation du nombre de sources à chaque fenêtre, BF_fixed_NbSrEstim+BSS 90

Table des figures

1.1 Analyse de scènes auditives par Romeo : étape de la séparation de sources 20
1.2 Le module de traitement audio du projet ROMEO 21
1.3 Le schéma d'annulation d'écho acoustique 22

2.1 Principe de la séparation de sources 30

3.1 Illustration d'un front d'onde plane arrivant sur un réseau de capteurs linéaire 45
3.2 Le cône de confusion 50
3.3 Les réponses impulsionnelles de têtes (HRIR : Head Related Impulse Response) : la représentation temporelle des HRTF pour différents micros de la tête du mannequin de vitrine 51
3.4 La réponse directionnelle $r_{\text{hrtf}}(f,\theta) = \mathbf{b}^H(f,\theta)\,\mathbf{d}_{\text{hrtf}}(f,\theta)$ (en dB) relative à différents angles de visée, construite à partir des HRTF . . 52

4.1 Principe de la séparation de sources 56
4.2 Parcimonie du signal de parole dans le domaine temps fréquence (b) comparé au domaine temporel (a) 60
4.3 Le paramètre p comme une fonction logistique, $p = p(t) = \frac{1}{1-exp(-L+\frac{(t-1)2L}{K_0})}$, L est le rang de calcul de la sigmoïde et $K_0 = 200$ est le nombre d'itération de l'algorithme 64

5.1 Principe de la séparation de sources avec un prétraitement de formation de voies 70
5.2 Formation de voies vers les directions d'arrivées 72
5.3 Formation de voies vers des directions de visée fixes 74
5.4 Formation de voies fixes avec sélection de lobes contenant les plus grandes énergies 75

5.5 Estimation du nombre de sources en utilisant la formation de voies fixe 76

6.1 Les fenêtres d'analyse utilisées dans la séparation adaptative des sources audio 83
6.2 Schéma d'adaptation pour l'algorithme de séparation de sources avec un prétraitement par formation de voies 85

7.1 Le réseau de capteurs de Theo 94
7.2 Paramètres extraits de la courbe de décroissance de l'énergie (EDC) (cette figure est extraite du manuel de cours d'Yves Grenier [36]) . . . 95
7.3 Réponse impulsionnelle entre un point dans une pièce et le microphone 1 96
7.4 Courbes de décroissance de l'énergie (EDC) entre un point dans une pièce et le premier microphone 97
7.5 Structure totale des séquences complémentaires de Golay à l'entrée et celles mesurées 100
7.6 Position des sources par rapport au réseau de capteurs dans le studio de Télécom ParisTech 101
7.7 Position des sources par rapport au réseau de capteurs dans l'Institut De la Vision (IDV) 102
7.8 Theo et la position des hauts-parleurs pour l'enregistrement des séquences complémentaires de Golay dans la chambre anéchoïque de Télécom ParisTech 103

9.1 Le rapport source-à-interférences SIR (courbes en bleu), le rapport source-à-distorsion SDR (courbes en rouge) et le rapport sources-à-artéfacts SAR (courbes en vert) des algorithmes BSS-l_1 (courbes continues) et ICA (courbes en tirets) : évaluation sur la base de données Theo-RI-studio. 114
9.2 Le score perceptuel global OPS (courbes en bleu), le score perceptuel relatif à la cible TPS (courbes en rouge), le score perceptuel relatif aux interférences IPS (courbes en vert) et le score perceptuel relatifs aux artéfacts APS (courbes en mauve) des algorithmes BSS-l_1 (courbes continues) et ICA (courbes en tirets) : évaluation sur la base de données Theo-RI-studio. 115
9.3 Les écarts types des résultats de séparations des algorithmes BSS-l_1 (barres grises) et ICA (barres violettes) : évaluation sur la base de données Theo-RI-studio. 115

9.4 Variation du rapport signal-à-interférences SIR de l'algorithme BSS-l_p en fonction du paramètre p : exemple de séparation de 4 paires de sources de la base de données Theo-RI-studio. 117
9.5 Le rapport source-à-interférences SIR (courbes en bleu), le rapport source-à-distorsion SDR (courbes en rouge) et le rapport sources-à-artéfacts SAR (courbes en vert) de l'algorithme BSS-l_p-param (courbes continues) en comparaison avec BSS-l_1 (courbes en tirets) : évaluation sur la base de données Theo-RI-studio. 117
9.6 Le score perceptuel global OPS (courbes en bleu), le score perceptuel relatif à la cible TPS (courbes en rouge), le score perceptuel relatif aux interférences IPS (courbes en vert) et le score perceptuel relatifs aux artéfacts APS (courbes en mauve) de l'algorithme BSS-l_p-param (courbes continues) en comparaison avec BSS-l_1 (courbes en tirets) : évaluation sur la base de données Theo-RI-studio. 118
9.7 Les écarts types des résultats de séparations des algorithmes BSS-l_p-param (barres grises) et BSS-l_1 (barres violettes) : évaluation sur la base de données Theo-RI-studio. 119
9.8 L'indice de Gini moyen de BSS-l_p-param au cours des itérations sur la base de données Theo-IR-Studio. 120
9.9 Evaluation du rapport source-à-interférences SIR de l'algorithme BSS-l_p-param au cours des itérations sur la base de données Theo-RI-Studio pour les différentes directions d'arrivées DOA considérées. 120
9.10 Influence du prétraitement par formation de voies en termes de rapport source-à-interférences SIR : évaluation sur la base de données Theo-RI-studio. 122
9.11 Influence du prétraitement par formation de voies en termes de rapport source-à-distorsion SDR : évaluation sur la base de données Theo-RI-studio. 123
9.12 Influence du prétraitement par formation de voies en termes de sources-à-artéfacts SAR : évaluation sur la base de données Theo-RI-studio. . 123
9.13 Influence du prétraitement par formation de voies en termes de score perceptuel global OPS : évaluation sur la base de données Theo-RI-studio. 124
9.14 Influence du prétraitement par formation de voies en termes de score perceptuel relatif à la source cible TPS : évaluation sur la base de données Theo-RI-studio. 124

9.15 Influence du prétraitement par formation de voies en termes de score perceptuel relatif aux interférences IPS : évaluation sur la base de données Theo-RI-studio. 125

9.16 Influence du prétraitement par formation de voies en termes de score perceptuel relatifs aux artéfacts APS : évaluation sur la base de données Theo-RI-studio. 125

9.17 Influence de l'angle inter-lobes de la formation de voies en termes de rapport source-à-interférences SIR : angle inter-lobes égal à 5° BF_fixed[5°]+BSS-l_1, 10° BF_fixed[10°]+BSS-l_1, 15° BF_fixed[15°]+BSS-l_1, et 30° BF_fixed[30°]+BSS-l_1, évaluation sur la base de données Theo-RI-studio. 126

9.18 Influence de l'angle inter-lobes de la formation de voies en termes de rapport source-à-distorsion SDR : angle inter-lobes égal à 5° BF_fixed[5°]+BSS-l_1, 10° BF_fixed[10°]+BSS-l_1, 15° BF_fixed[15°]+BSS-l_1, et 30° BF_fixed[30°]+BSS-l_1, évaluation sur la base de données Theo-RI-studio. 127

9.19 Influence de l'angle inter-lobes de la formation de voies en termes de rapport sources-à-artéfacts SAR : angle inter-lobes égal à 5° BF_fixed[5°]+BSS-l_1, 10° BF_fixed[10°]+BSS-l_1, 15° BF_fixed[15°]+BSS-l_1, et 30° BF_fixed[30°]+BSS-l_1, évaluation sur la base de données Theo-RI-studio. 127

9.20 Influence de l'angle inter-lobes de la formation de voies en termes de score perceptuel global OPS : angle inter-lobes égal à 5° BF_fixed[5°]+BSS-l_1, 10° BF_fixed[10°]+BSS-l_1, 15° BF_fixed[15°]+BSS-l_1, et 30° BF_fixed[30°]+BSS-l_1, évaluation sur la base de données Theo-RI-studio. 128

9.21 Influence de l'angle inter-lobes de la formation de voies en termes de score perceptuel relatif à la cible TPS : angle inter-lobes égal à 5° BF_fixed[5°]+BSS-l_1, 10° BF_fixed[10°]+BSS-l_1, 15° BF_fixed[15°]+BSS-l_1, et 30° BF_fixed[30°]+BSS-l_1, évaluation sur la base de données Theo-RI-studio. 128

9.22 Influence de l'angle inter-lobes de la formation de voies en termes de score perceptuel relatif aux interférences IPS : angle inter-lobes égal à 5° BF_fixed[5°]+BSS-l_1, 10° BF_fixed[10°]+BSS-l_1, 15° BF_fixed[15°]+BSS-l_1, et 30° BF_fixed[30°]+BSS-l_1, évaluation sur la base de données Theo-RI-studio. 129

9.23 Influence de l'angle inter-lobes de la formation de voies en termes de score perceptuel relatifs aux artéfacts APS : angle inter-lobes égal à 5° BF_fixed[5°]+BSS-l_1, 10° BF_fixed[10°]+BSS-l_1, 15° BF_fixed[15°]+BSS-l_1, et 30° BF_fixed[30°]+BSS-l_1, évaluation sur la base de données Theo-RI-studio. 129

9.24 Influence de la sélection de lobes en termes de rapport source-à-interférences SIR : évaluation sur la base de données Theo-RI-studio. 131

9.25 Influence de la sélection de lobes en termes de rapport source-à-distorsion SDR : évaluation sur la base de données Theo-RI-studio. 132

9.26 Influence de la sélection de lobes en termes de rapport sources-à-artéfacts SAR : évaluation sur la base de données Theo-RI-studio. . . . 132

9.27 Influence de la sélection de lobes en termes de score perceptuel global OPS : évaluation sur la base de données Theo-RI-studio. 133

9.28 Influence de la sélection de lobes en termes de score perceptuel relatif à la cible TPS : évaluation sur la base de données Theo-RI-studio. . . 133

9.29 Influence de la sélection de lobes en termes de score perceptuel relatif aux interférences IPS : évaluation sur la base de données Theo-RI-studio.134

9.30 Influence de la sélection de lobes en termes de score perceptuel relatifs aux artéfacts APS : évaluation sur la base de données Theo-RI-studio. 134

9.31 Les écarts types des résultats de séparations des algorithmes BSS-l_1 (barres grises), BF_fixed[5°]+BSS-l_1 (barres violettes), BF_DOA+BSS-l_1 (barres vertes) et BF_fixed[5°]_BS+BSS-l_1 (barres rouges) : évaluation sur la base de données Theo-RI-studio. 135

9.32 Estimation des directions d'arrivées en utilisant l'algorithme de formation de voies BF_fixed[5°]_BS pour les 40 cas de séparation de source de la base de données Theo-RI-studio. 135

9.33 Étude de la vitesse de convergence : la valeur de la fonction de coût de (a) BSS-l_1 et (b) BF_fixed[5°]_BS+BSS-l_1 à travers les itérations et pour différentes directions d'arrivées. 136

9.34 Une vue de dessus des différentes configuration du réseau de capteurs : le nombre de microphones varie de 2 à 16. 137

9.35 Variation du rapport source-à-interférences SIR (courbes en bleu), le rapport source-à-distorsion SDR (courbes en rouge) et le rapport sources-à-artéfacts SAR (courbes en vert) avec le nombre de capteurs pour la séparation de deux sources à partir de différentes directions d'arrivées : évaluation de BF_fixed[5°]_BS+BSS-l_1 et BSS-l_1 : évaluation sur la base de données Theo-RI-studio. 139

9.36 Variation des mesures perceptuelles avec le nombre de capteurs pour la séparation de deux sources à partir de différentes directions d'arrivées : évaluation de BF_fixed[5°]_BS+BSS-l_1 : évaluation sur la base de données Theo-RI-studio. 139

9.37 Variation du rapport source-à-interférences SIR avec le nombre de capteurs et pour différentes directions d'arrivées pour l'algorithme BF_fixed[5°]_BS+BSS-l_1 : évaluation sur la base de données Theo-RI-studio. 140

9.38 Variation du rapport source-à-distorsion SDR avec le nombre de capteurs et pour différentes directions d'arrivées pour l'algorithme BF_fixed[5°]_BS+BSS-l_1 : évaluation sur la base de données Theo-RI-studio. 140

9.39 Variation du rapport sources-à-artéfacts SAR avec le nombre de capteurs et pour différentes directions d'arrivées pour l'algorithme BF_fixed[5°]_BS+BSS-l_1 : évaluation sur la base de données Theo-RI-studio. 141

9.40 Variation du score perceptuel global OPS avec le nombre de capteurs et pour différentes directions d'arrivées pour l'algorithme BF_fixed[5°]_BS+BSS-l_1 : évaluation sur la base de données Theo-RI-studio. 141

9.41 Variation du score perceptuel relatif à la cible TPS avec le nombre de capteurs et pour différentes directions d'arrivées pour l'algorithme BF_fixed[5°]_BS+BSS-l_1 : évaluation sur la base de données Theo-RI-studio. 142

9.42 Variation du score perceptuel relatif aux interférences IPS avec le nombre de capteurs et pour différentes directions d'arrivées pour l'algorithme BF_fixed[5°]_BS+BSS-l_1 : évaluation sur la base de données Theo-RI-studio. 142

9.43 Variation du score perceptuel relatif aux artéfacts APS avec le nombre de capteurs et pour différentes directions d'arrivées pour l'algorithme BF_fixed[5°]_BS+BSS-l_1 : évaluation sur la base de données Theo-RI-studio. 143

9.44 Variation du rapport source-à interférences SIR et du rapport source-à-distorsion SDR de l'algorithme BF_fixed[5°]_BS+BSS-l_1 avec le nombre de capteurs : évaluation sur la base de données Theo-RI-IDV avec deux longueur de réponses impulsionnelles : RI1 = 500 échantillons et RI2 = 800 échantillons. 143

9.45 La moyenne de la norme l_1 au cours des itérations de l'algorithme BF_fixed[5°]_BS+BSS-l_1 pour différentes tailles du réseau de capteurs : évaluation sur la base de données Theo-RI-studio. 144

10.1 Exemple d'un cas de séparation avant le mélange (sources bruts) avec un nombre de sources variant entre 1 et 2. 148

10.2 Le rapport source-à-interférences SIR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse longues : nombre de sources réel fixe et connu (2 sources). 149

10.3 Le rapport source-à-distorsion SDR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse longue : nombre de sources réel fixe et connu (2 sources). 150

10.4 Le rapport sources-à-artéfacts SAR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse longues : nombre de sources réel fixe et connu (2 sources). 150

10.5 Le rapport source-à-interférences SIR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse longues : nombre de sources réel connu et variable entre 1 et 2 sources. 151

10.6 Le rapport source-à-distorsion SDR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse longues : nombre de sources réel connu et variable entre 1 et 2 sources. 151

10.7 Le rapport source-à-artéfacts SAR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse longues : nombre de sources réel connu et variable entre 1 et 2 sources. 152

10.8 Le rapport source-à-interférences SIR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse : nombre de sources réel est inconnu et fixé *a priori*. 153

10.9 Le rapport source-à-distorsion SDR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse : nombre de sources réel est inconnu et fixé *a priori*. 153

10.10Le rapport sources-à-artéfacts SAR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse : nombre de sources réel connu et variable entre 1 et 2 sources. 154
10.11Le rapport source-à-interférences SIR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse : le nombre de sources réel est inconnu et fixé *a priori*. 154
10.12Le rapport source-à-distorsion SDR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse : le nombre de sources réel est inconnu et fixé *a priori*. 155
10.13Le rapport source-à-artéfacts SAR de BF_fixed[5°]_BS+BSS-l_1 au cours des fenêtres d'analyse : le nombre de sources réel est inconnu et fixé *a priori*. 155
10.14Estimation du nombre de source au cours des fenêtres d'analyse avec BF_fixed[5°]_NbSrEstim : le nombre réel de sources est fixe. 156
10.15Estimation du nombre de source au cours des fenêtres d'analyse avec BF_fixed_NbSrEstim : le nombre réel de sources varie entre 1 et 2. . 157
10.16Le rapport source-à-interférences SIR de BF_fixed[5°]_NbSrEstim+BSS-l_1 au cours des fenêtres d'analyse : le nombre de sources est estimé (nombre de sources réel égal à 2). 158
10.17Le rapport source-à-distorsion SDR de BF_fixed[5°]_NbSrEstim+BSS-l_1 au cours des fenêtres d'analyse : le nombre de sources est estimé (nombre de sources réel égal à 2). 158
10.18Le rapport sources-à-artéfacts SAR de BF_fixed[5°]_NbSrEstim+BSS-l_1 au cours des fenêtres d'analyse : le nombre de sources est estimé (nombre de sources réel égal à 2). 159
10.19Le rapport source-à-interférences SIR de BF_fixed[5°]_NbSrEstim+BSS-l_1 au cours des fenêtres d'analyse : le nombre de sources est estimé et variable entre 1 et 2 sources. 159
10.20Le rapport source-à-distorsion SDR de BF_fixed[5°]_NbSrEstim+BSS-l_1 au cours des fenêtres d'analyse : le nombre de sources est estimé et variable entre 1 et 2 sources. 160
10.21Le rapport sources-à-artéfacts SAR de BF_fixed[5°]_NbSrEstim+BSS-l_1 au cours des fenêtres d'analyse : le nombre de sources est estimé et variable entre 1 et 2 sources. 161
10.22Le rapport source-à-interférences SIR moyen sur toutes les fenêtres d'analyse longues : nombre de sources fixe au cours des itérations. . . 162

10.23Le rapport source-à-distorsion SDR moyen sur toutes les fenêtres d'analyse longues : nombre de sources fixe au cours des itérations. . . 162
10.24Le rapport sources-à-artéfacts SAR moyen sur toutes les fenêtres d'analyse longues : nombre de sources fixe au cours des itérations. . . 163
10.25Le rapport source-à-interférences SIR moyen sur toutes les fenêtres d'analyse longues : nombre de sources variable au cours des itérations 163
10.26Le rapport source-à-distorsion SDR moyen sur toutes les fenêtres d'analyse longues : nombre de sources variable au cours des itérations 164
10.27Le rapport sources-à-artéfacts SAR moyen sur toutes les fenêtres d'analyse longues : nombre de sources variable au cours des itérations 164
10.28Le rapport source-à-interférences (SIR), le raport source-à-distorsion (SAR) et le rapport sources-à-artéfacts (SAR) des sources estimées avec HARK (en bleu) et BF_fixed[5°]_NbSrEstim+BSS-l_1 (en vert) et leurs écarts types (barres verticales) : les moyennes pour la séparation de 40 paires de sources de la base de données Theo-RI-studio, les directions d'arrivées sont 0° et 60°. 168
10.29Le score perceptuel global (OPS), le score perceptuel relatif à la cible (TPS), le score perceptuel relatif aux interférences (IPS) et le score perceptuel relatif aux artéfacts (APS) des sources estimées avec HARK (en bleu) et BF_fixed[5°]_NbSrEstim+BSS-l_1 (en vert) et leurs écarts types (barres verticales) : les moyennes pour la séparation de 40 paires de sources de la base de données Theo-RI-studio, les directions d'arrivées sont 0° et 60°. 169

11.1 Romeo 173
11.2 La position des capteurs dans la tête de Romeo. 173

Index

Analyse computationnelle de scènes auditives, 18
Analyse de scènes auditives, 18
Analyse en composantes indépendantes (ICA), 57
Audition des robots, 18

Cocktail party, 18
Courbe de décroissance de l'énergie, 95

Diagramme de directivité, 43
Différence d'intensité interaural (IID), 49
Différence de temps interaural (ITD), 49
Direction d'arrivée, 44

Filtrage spatio-temporel, 44
Fonction de transfert de tête (HRTF), 24, 50
Formation de voies, 43
Formation de voies adaptative, 44
Formation de voies fixe, 44

Gradient naturel, 56

Information mutuelle, 58

Mélanges convolutifs, 19

Projet Romeo, 17

Réponse directionnelle, 46
Réponse impulsionnelle acoustique, 94
Réponse impulsionnelle de tête (HRIR), 50

Séparation aveugle de sources, 19
Séparation de sources, 19
Séparation de sources déterminée, 20
Séparation de sources sous-déterminée, 20

Vecteur directionnel, 46

Séparation de sources pour l'audition des robots

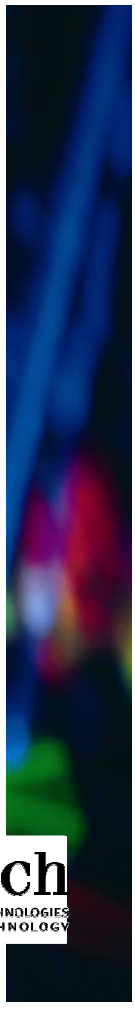
RESUME : Cette thèse propose des algorithmes de séparation aveugle de sources audio en utilisant un réseau de capteurs. L'application finale de ces algorithmes est l'audition des robots dans le cadre du projet ROMEO. Dans cette thèse, nous avons développé des algorithmes de séparation aveugle de sources audio basés sur des critères de parcimonie. Nous montrons que la minimisation de la norme l_1 avec une technique d'optimisation du gradient naturel permet d'élaborer un algorithme se situant au niveau de l'état de l'art. Nous montrons qu'un critère basé sur la paramétrisation de la pseudo-norme l_p, avec $0<p<1$ en améliore les performances. Ceci revient à rendre la contrainte de parcimonie plus dure au fur et à mesure que l'algorithme avance dans ses itérations. Pour exploiter l'aspect multicapteurs de notre application (16 capteurs sont fixés autour de la tête de l'humanoïde), nous avons proposé des algorithmes de séparation avec une étape de prétraitement de formation de voies fixe. Dans le cas de l'audition des robots, les capteurs sont souvent placés sur la tête de l'humanoïde. Afin de tenir compte de l'influence de la tête sur le champ sonore proche, nous avons construit les filtres de formation en utilisant les fonctions de transfert de tête (HRTF) du robot. L'étape de formation de voies améliore les résultats de séparation par rapport à l'utilisation d'un algorithme de séparation seule de minimum 10dB. Cette thèse propose aussi les versions adaptatives de ces algorithmes. Dans le scénario réel où le nombre de sources présentes dans l'environnement du robot est inconnu et change au cours du temps, nous montrons comment détecter et suivre le nombre de sources.

Mots clés : Séparation aveugle de sources, réseaux de capteurs, formation de voies, HRTF

Source separation for robot audition

ABSTRACT : This thesis proposes blind audio source separation algorithms using a microphone array. The final application of these algorithms is robot audition through the ROMEO project. In this thesis, we developed blind source separation algorithms based on a sparcity criterion. We show that l_1 minimization using the natural gradient optimization technique has the same performance that the state of the art. We show that a criterion based on the parametrization of the quazi-norm l_p, with $0<p<1$, improves the previous results: the sparsity criterion gets harder through the iterations of the algorithm. Then, we exploited the multisensor aspect of our application (16 sensors are fixed in the head of the humanoid) and we proposed a separation algorithms with a fixed beamforming preprocessing step. In the robot audition case, the sensors are often placed on the head of the humanoid. To take into account the influence of the head in the near sound manifold, we built the beamforming filters using the head related transfer functions (HRTF) of the robot. The beamforming step improves the separation results compared to the use of a blind source separation only. This thesis also proposes the adaptive versions of those algorithms. In the real scenario where the number of sources is unknown and changes, we show how to detect and follow the number of sources.

Keywords : Blind source separation, microphone array, beamforming, HRTF



TELECOM ParisTech

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY