Contexte général Pro jet 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).

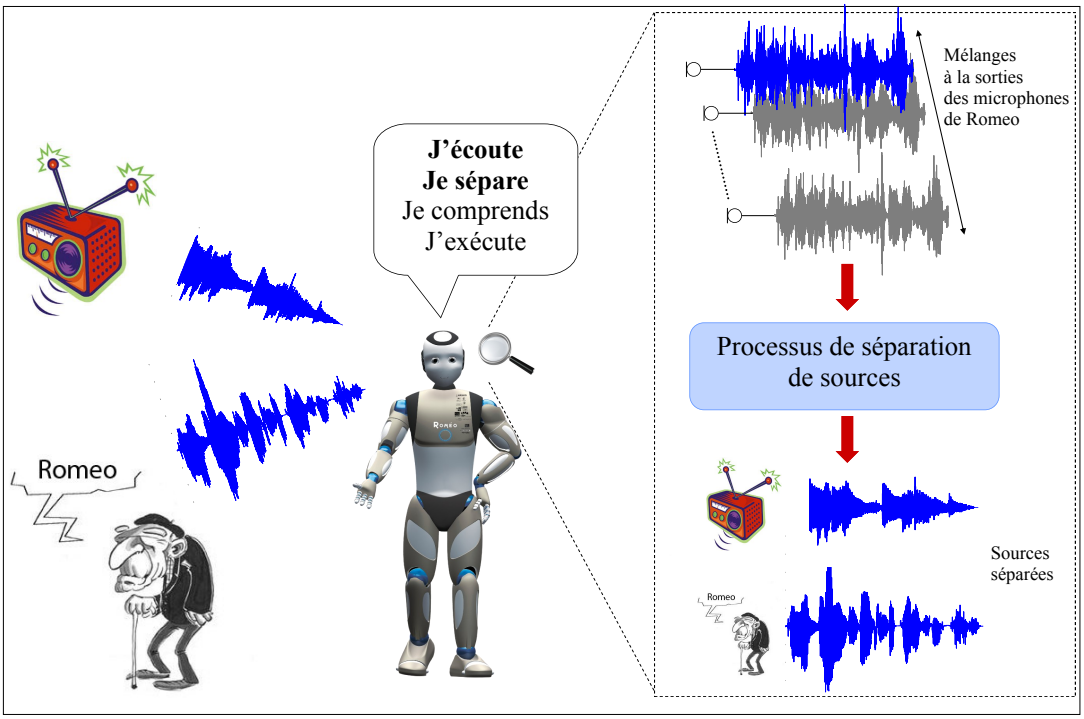

Figure 1.1 – Analyse de scènes auditives par Romeo : étape de la séparation de sources

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) :

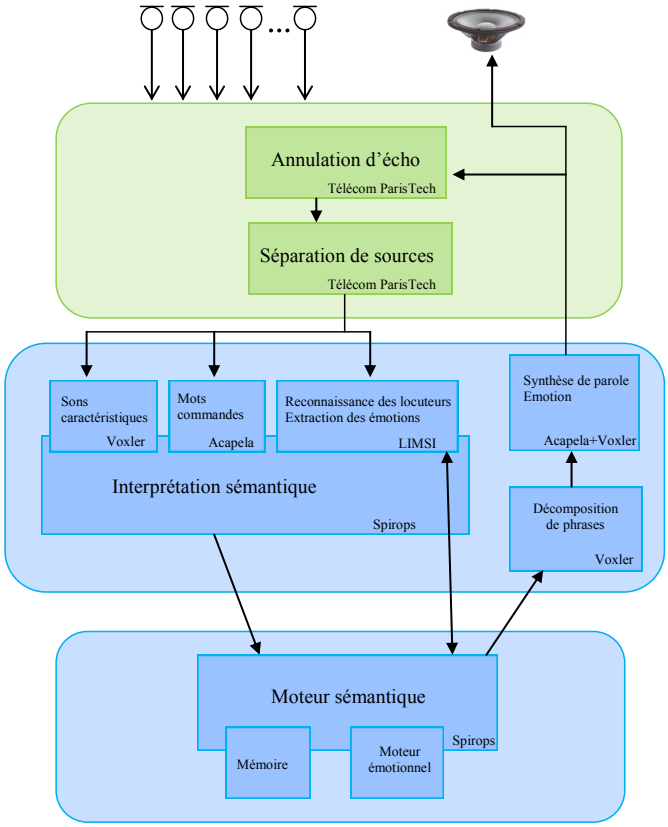

Figure 1.2 – Le module de traitement audio du projet Romeo

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.

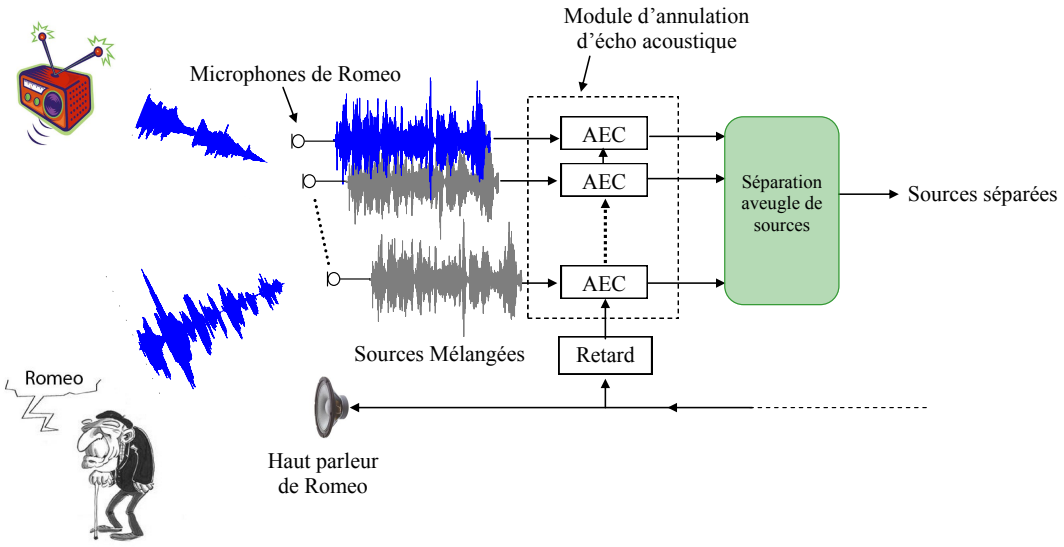

FIGURE 1.3 – Le schéma d'annulation d'écho acoustique

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.