Un (petit) pas vers la perception interactive

Sylvain Argentieri

SORBONNE UNIVERSITÉ

HABILITATION À DIRIGER DES RECHERCHES

par

Sylvain ARGENTIERI

Un (petit) pas vers la perception interactive

Laboratoire : Institut des Systèmes Intelligents et de Robotique, UMR UPMC/CNRS 7222

Soutenue le 06/12/2018 devant le jury composé de :

David FILLIAT	Professeur à l'ENSTA Paritech	Rapporteur
François MICHAUD	Professeur à l'Université de Sherbrooke, Canada	Rapporteur
Mathias QUOY	Professeur à l'Université de Cergy-Pontoise	Rapporteur
Philippe BIDAUD	Professeur à Sorbonne Université	Examinateur
Philippe SOUÈRES	Directeur de Recherche au LAAS-CNRS	Examinateur
Bruno GAS	Professeur à Sorbonne Université	Référent HDR

« *Les grandes personnes ne comprennent jamais rien toutes seules, et c'est fatiguant, pour les enfants, de toujours et toujours leur donner des explications. Pourtant, toutes les grandes personnes ont d'abord été des enfants. (Mais peu d'entre elles s'en souviennent.)* »

Adapté de Antoine de Saint-Exupéry

Table des matières

Chapitre 1

Introduction

Perception (nom féminin, du latin perceptio, -onis)

— Action de percevoir par les organes des sens : La perception des couleurs.

— Événement cognitif dans lequel un stimulus ou un objet, présent dans l'environnement immédiat d'un individu, lui est représenté dans son activité psychologique interne, en principe de façon consciente ; fonction psychologique qui assure ces perceptions.

Larousse

On trouve trace très tôt dans l'histoire de la volonté des Hommes de construire des machines à même d'opérer des tâches répétitives. Aujourd'hui qualifiées d'automates (MEYER, 2015), elles visaient également à tenter de reproduire (et donc comprendre) les mécanismes du mouvement, que ce soit chez l'Homme ou les animaux. Au 18ème siècle, on venait ainsi voir ces machines curieuses reproduire certaines capacités du vivant, comme a pu le faire le fameux canard de Jacques de Vaucanson (HEUDIN, 2008), capable de se nourrir, de caqueter et digérer sa nourriture. Ce n'est qu'en 1920 que le terme robot apparaît, à l'occasion d'une pièce de théâtre où sont ainsi désignés des androïdes à l'apparence humaine. Les premiers robots ainsi qualifiés apparaissent dans la foulée, souvent inspirés d'animaux (chien électrique de Hammond et Miessner, tortues cybernétiques de W. Grey Walter, etc.). Ces premières machines se distinguent des automates précédents par leur capacité à réagir à leur environnement. Elles sont ainsi équipées d'organes sensoriels (des capteurs) qui viennent influencer l'activité de leurs organes moteurs. S'en suivi l'apparition des premiers robot industriels, ainsi que la création des premiers laboratoires dédiés à leur étude : la Robotique, en tant que discipline, était née. Le terme "robotique" fait aujourd'hui parti de notre quotidien. Si ce n'est probablement pas la définition qu'en donnerait le grand public, on peut néanmoins tenter d'en préciser les contours d'une manière (très) simplifiée. Un robot peut être vu comme la réunion :

— de capteurs qui servent à renseigner le robot sur son environnement et sur son état interne ; ces éléments s'apparentent aux sens, comme par exemple la vision, l'audition, etc. Ces capacités ne sont pas nécessairement limitées à celles de l'Homme : un robot peut être sensible aux ultrasons, voir dans l'infrarouge, etc. ;

— d'actionneurs permettant l'interaction mécanique du robot avec ce qui l'entoure via le déplacement de parties de son propre corps, ou au sein même de son environnement ;

— de capacités de traitement de ces informations, de raisonnement et de décision, souvent implémentées logiciellement au sein de calculateurs numériques embarqués ou déportés.

Et c'est l'union indissociable de ces trois éléments qui fournit à un robot la capacité d'analyser son environnement, de décider et d'agir en son sein. A cette définition historique on pourrait ajouter aujourd'hui les capacités d'autonomie et d'adaptation dont il semble maintenant difficile de faire abstraction dans les applications récentes de la robotique, comme

l'assistance aux personnes ou les applications plus industrielles autour de l'agriculture et les transports par exemple.

Contexte

J'ai eu l'occasion de travailler ces dix dernières années sur une de ces trois briques de base qui semblent définir la robotique : la perception. Cette perception a longtemps été envisagée selon une définition proche de celle suggérée par la définition du Larousse et rappelée en début de cette introduction : la capacité à interpréter les données issues des organes des sens. Cette interprétation peut s'effectuer à un niveau très bas, directement en sortie des senseurs : par exemple, la perception de la couleur rouge est immédiatement accessible en sortie de notre œil, grâce aux photorécepteurs dont il est équipé. A vrai dire, il serait bien plus juste de parler ici de sensation de rouge. Car cette même perception peut s'effectuer à un niveau plus haut, peut être plus cognitif : cette même couleur est alors interprétée, ressentie, comme rouge et non pas comme une autre couleur. Pour continuer cette illustration avec la modalité visuelle, on peut aussi se demander comment *cette catastrophe qu'est l'œil* –comme s'amuse à le dénommer K. O'Regan dans (J. O'REGAN, 2011)– nous donne accès de manière consciente à une scène stable, nette et complète alors que tout dans la physiologie de l'œil et son fonctionnement s'oppose à ces ressentis qualitatifs.

Ces questions fondamentales semblent encore loin de la façon dont la perception est abordée dans un contexte robotique, où une distinction est traditionnellement faite entre la proprioception –c'est à dire la capacité pour un robot de ressentir son état interne (en termes de positions ou vitesses angulaires, d'orientation, etc.)– et l'extéroception –qui renseigne le robot sur l'état de son environnement (via des ultrasons, des caméras, des lasers, des microphones, etc.)–. Au sein de cette dernière, la vision est probablement la modalité la plus utilisée au sein des tâches robotiques les plus variées. La richesse des informations visuelles et le coût devenu négligeables de caméras hautes performances explique probablement cet état de fait. A bien y regarder, les autres modalités n'ont pas reçu autant d'attention par la Communauté Robotique. En particulier, et bien que l'audition soit reconnue comme un sens critique à l'interaction et la socialisation entre humains, cette modalité n'aura été identifiée comme une thématique scientifique pertinente dans un contexte robotique qu'au début des années 2000. C'est précisément sur cette thématique que j'ai effectué ma thèse, soutenue en 2006, et sur laquelle j'ai eu l'occasion de continuer à travailler au sein de l'Institut des Systèmes Intelligents et de Robotique (ISIR) depuis 2008. Doter un robot de capacités de perception sonore est séduisant ; cela permet en particulier d'envisager un des moyens de communication les plus naturels de l'homme : sa voix. Mais cela complémente également de manière pertinente les informations visuelles éventuellement disponibles. J'aurai l'occasion d'illustrer le rôle important de cette multimodalité dans la première partie de ce document.

Cependant, doter une machine de capacités de perception seules n'en fait pas un robot. Comme énoncé précédemment, l'action de cette machine sur et dans son environnement est au moins nécessaire. Et alors que beaucoup de travaux en Robotique envisage le processus de perception comme passif et instantané, les contributions synthétisées dans ce manuscrit envisagent l'existence de l'action parfois comme une contrainte, mais surtout comme une opportunité pour repenser le sens de cette perception. Ainsi, le rôle de l'action ne se retrouve plus disjoint de celui de la perception, et ces deux capacités peuvent toutes deux être entremêlées au point qu'il semble difficile d'aborder l'une sans l'autre. C'est néanmoins sous le prisme de la perception, vue dans ce document comme la capacité pour un robot à comprendre son interaction avec son environnement, que j'ai abordé cette problématique. Il faut bien comprendre qu'il ne s'agit pas de remettre en cause tous les travaux déjà effectués dans le domaine de la perception en Robotique : ils sont nombreux, impressionnants, et les résultats obtenus permettent aujourd'hui de résoudre des problématiques très concrètes. On sait aujourd'hui exploiter des données issues de capteurs très différents pour rendre une

voiture autonome ; reconnaître une image ou un son, même des des situations difficiles ; faire apparaître des objets virtuels au sein d'un environnement réel ; etc. L'enjeu ici est plutôt de tenter de proposer un cadre plus général dans lequel peut être exploité la perception et l'action afin d'atteindre ce quatrième objectif qu'est l'autonomie. J'aurai également l'occasion de discuter de ces enjeux plus en détails dans la seconde partie de ce document.

Structure du document

Le document est structuré de la façon suivante. Après cette courte introduction, un premier chapitre est dédié à mes contributions à l'analyse de scène sonore en Robotique. Après quelques éléments de contextualisation présentés dans une première partie, les deux parties suivantes sont dédiées aux travaux que j'ai mené autour de la localisation binaurale de source sonore. Enfin, des considérations actives sont introduites dans une dernière partie, où la modulation du mouvement de tête d'un robot équipé de capacités audio et visuelles sera abordée.

Je présente alors dans un second chapitre mes contributions pour une approche interactive de la perception. Radicalement différent dans son approche de mes contributions précédentes, ce chapitre est l'occasion pour moi d'introduire un point de vue sensorimoteur de la perception, au travers de l'estimation de la dimension de l'espace et de l'extraction de structures invariantes au sein du flux sensorimoteur.

Enfin, une conclusion termine ce document. J'en profite pour esquisser quelques pistes sur mon travail futur.

Tout au long du document, le lecteur trouvera des boites de texte grisées. Elles sont l'occasion pour moi de préciser le contexte et la direction des travaux ayant donnés lieux aux résultats synthétisés ensuite.

Publication

Ces parties sont en général conclues par une sélection des publication (en journal ou conférence) effectuées par l'étudiant que j'ai co-encadré.

Les travaux synthétisés dans ce manuscrit ont été menés par les étudiants suivants, durant la période précisée (le taux d'encadrement, ainsi que leur directeur de thèse sont également mentionnés) :

- Alban LAFLAQUIÈRE (septembre 2009 - juin 2013), co-encadré à 50% avec Bruno GAS ;
- Alban PORTELLO (septembre 2010 - décembre 2013), co-encadrée à 20% avec Patrick DANÈS ;
- Karim YOUSSEF (septembre 2010 - octobre 2013), co-encadré à 80% avec Jean-Luc ZARADER ;
- Benjamin COHEN-LHYVER (décembre 2013 - septembre 2017), co-encadré à 80% avec Bruno GAS ;
- Valentin MARCEL (septembre 2015 - *encore en cours*), co-encadré à 80% avec Bruno GAS.

Si ce document a vocation à présenter le contexte général de mon travail d'encadrement, c'est bien en collaboration avec ces étudiants (et leur directeur de thèse) que les travaux présentés dans la suite ont été effectués. A ce titre, j'emploierai dans ce manuscrit non pas un "nous" de politesse, mais bien un "nous" collectif qui reflétera j'espère ce travail commun que j'ai eu la chance de partager avec toutes ces personnes.

Chapitre 2

Contributions à l'analyse de scène sonore en Robotique

Sommaire

Selon (BERGMAN, 1990), l'analyse de scène **auditive** (ASA) désigne habituellement l'ensemble des processus grâce auxquels notre système auditif transforme un mélange de sons, en provenance d'un environnement acoustique complexe, en des entités perceptives indépendantes (i.e. des *objets sonores*, spatialisés ou non). Son pendant computationnel, connu sous le terme d'analyse computationnelle de scène auditive (CASA), est alors exploité dans le domaine de la perception artificielle pour doter des systèmes de capacités d'analyse automatique d'une scène sonore. Néanmoins, la définition même de l'ASA met l'homme et ses

capacités d'interprétation au cœur de la problématique de l'analyse de la scène. En d'autres termes, l'ASA comporte une partie importante de psychoacoustique, i.e. d'interprétations subjectives d'une scène acoustique. Et c'est d'ailleurs un de ses objectifs que de comprendre, interpréter, modéliser et reproduire dans un cadre computationnel les formidables capacités de notre système auditif. A l'opposé, on a longtemps attendu des systèmes de perception artificielle, et en particulier dans un contexte robotique, qu'ils soient capables de proposer une analyse *objective* de la scène acoustique. Le système d'analyse automatique doit alors être capable d'estimer les paramètre *physiques* de la scène acoustique (position, intensité, hauteur, etc.) des sources acoustiques. On compare alors ses performances par rapport à la vérité terrain, et l'objectif est alors que l'analyse de la scène **sonore** (ASS) effectuée se rapproche au plus prêt de sa description acoustique physique. Cette distinction entre analyse de scène auditive et sonore est rarement effectuée dans la littérature. Très souvent, et nous le ferons nous même largement dans la suite de ce manuscrit, l'ASA est le terme consacré pour qui s'intéresse à la perception artificielle des sons. Pourtant, cette distinction est la clé pour comprendre en quoi et comment le contexte robotique est à même de proposer des solutions originales au fameux *cocktail party problem* (BRONKHORST, 2000), et plus spécifiquement dans un cadre binaural.

Nous proposons de présenter dans ce chapitre nos différentes contributions dans le domaine de l'analyse de scène sonore en Robotique. De fait, aucune (ou très peu) de considérations psychoacoustique entreront en jeu dans la suite. Ce chapitre est composé comme suit. Dans une première section nous préciserons le contexte de ces travaux, en mettant en avant le contexte robotique et ses spécificités. Puis dans les sections suivantes nous détaillerons différents travaux, portant autant sur les aspects dits « bas-niveau » (extraction de caractéristiques, localisation) que plus « haut-niveau » (modulation attentionnelle du mouvement de la tête). Comme indiqué dans l'introduction de ce manuscrit, une attention particulière sera portée à l'aspect actif ou interactif des méthodes.

2.1 Introduction

2.1.1 Contexte

Une célèbre citation, attribuée à Hellen Keller [1] nous rappelle que "la cécité sépare les gens des objets. La surdité sépare les gens les uns des autres" (KOHLRAUSCH et al., 2013). L'audition est en effet une modalité critique à l'interaction et la socialisation. Pour autant, s'il apparaît difficile aujourd'hui de se passer d'informations auditives, son exploitation dans un contexte robotique n'a attiré l'attention que tardivement, du moins comparativement à la modalité visuelle. L'explosion des problématiques liées à l'interaction homme-machine a sans nul doute contribué, au début des années 2000, à la redécouverte de cette modalité. C'est ainsi que la communauté "*Robot Audition*" a émergé, dans un contexte où il était encore difficile d'imaginer des systèmes auditifs artificiels autres que bio-inspirés (NAKADAI, LOURENS et al., 2000). Pour autant, ce nouveau contexte (robotique) interroge de manière pertinente les approches déjà nombreuses d'analyse d'une scène sonore au regard des différentes contraintes qu'il apporte. Parmi celles-ci, on retrouve (ARGENTIERI, DANÈS et SOUÈRES, 2015) :

— contrainte géométrique : le système auditif artificiel doit pouvoir être embarqué sur une plateforme robotique, humanoïde ou non. Deux philosophies s'affrontent ici :
  - l'approche "antennerie" : le capteur audio est ici constitué d'un réseau de microphones, et la redondance de ce type de design permet d'envisager tout type

1. H. Keller (1880–1968) est la première femme aveugle et sourde à avoir obtenu un diplôme universitaire aux États-Unis. Auteur et activiste engagée, elle a écrit de nombreux ouvrage sur ses convictions et sa condition.

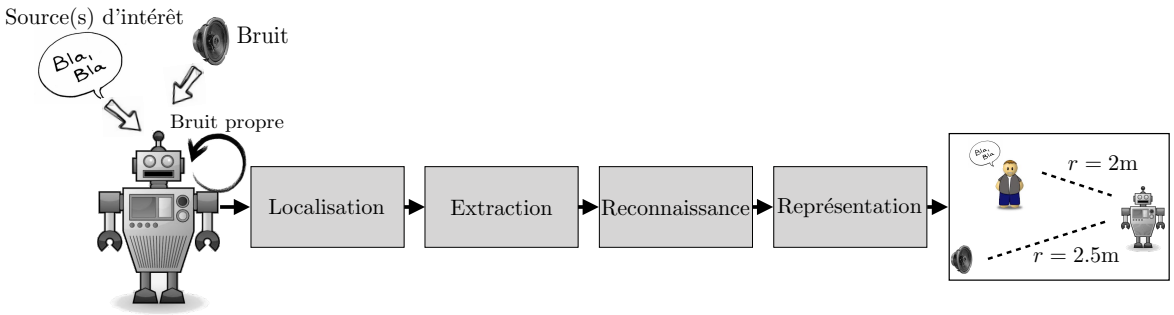


FIGURE 2.1 – Organisation *bottom-up* traditionnelle, depuis le signal audio récupéré sur le robot à son interprétation. Reproduction de (ARGENTIERI, PORTELLO et al., 2013).

d'analyse de scène sonore. C'est aujourd'hui le type de capteur le plus utilisé en robotique, car les performances obtenues sont très bonnes. La littérature sur les antennes de microphones est aussi très riche (VAN TREES, 2002 ; BENESTY, CHEN et HUANG, 2008), ces capteurs étant depuis longtemps utilisés en traitement du signal et acoustique ;

- l'approche "binaurale" : le capteur audio est ici constitué de seulement deux microphones, placés généralement de part et d'autre d'une "tête" robotique dont l'effet acoustique est primordial à la mise en place des algorithmes dits "binauraux" d'analyse audio. Ce type de design convient particulièrement aux robot humanoïdes, sans être exclusifs (LÖLLMANN et al., 2017). De même, l'aspect bio-inspiré est logiquement présent pour ce type d'approche, bien que les traitements effectués en sorties des microphones ne le soient pas nécessairement. Ce contexte binaural reste privilégié pour qui souhaite s'inspirer, reproduire et évaluer des modèles d'audition humaine.

— contrainte temporelle : l'exploitation de données audio doit pouvoir se faire de manière réactive dans un contexte robotique incluant souvent des comportements réflexes. Analyser une scène audio nécessite donc une rapidité de traitement importante, que peu d'approches déjà existantes peuvent revendiquer. Un certain nombre de travaux ont ainsi eu pour objectif d'optimiser autant que possibles les temps de calculs, tandis que d'autres contributions ont proposé des solutions matérielles (LUNATI, MANHÈS et DANÈS, 2012) (via des systèmes hardware dédiés) et/ou logiciels (NAKADAI, OKUNO et al., 2008 ; GRONDIN et al., 2013) (modules spécifiques de middlewares robotiques ou framework à part entière) à cette contrainte ;

— contrainte environnementale : les applications robotiques modernes envisagent des environnements fondamentalement dynamiques et hautement imprédictibles. Cela se traduit par des conditions acoustiques variables, incluant bruits et réverbérations auxquels les méthodes devront être robustes. Une originalité du contexte robotique réside également dans le fait que le robot lui-même participe à ces perturbations : les bruits propres, issus d'actionneurs ou des ventilateurs sont souvent à l'origine de bruits importants venant dégrader les performances de l'analyse (INCE, 2011 ; OKUTANI et al., 2012).

Pour tenter de résoudre tout ou partie de ces problématiques, l'essentiel des travaux existant s'appuie sur une architecture classique dite *bottom-up*, partant du signal "brut" pour aller vers une représentation plus haut-niveau, très souvent traduite sous la forme d'une carte spatiale de l'environnement du robot, cf. figure 2.1. Dans ce type d'architecture, les signaux issus des capteurs sonores sont très souvent d'abord exploités pour *localiser* la ou les sources

sonores présentes dans l'environnement. Il s'agit ici d'estimer, en terme spatial (distance, azimut, élévation par exemple), les positions relatives des sources sonores présentes autour du robot. Sur la base de ces positions estimées, des algorithmes de séparation de source sont alors mis en œuvre pour *extraire* de la mixtures de signaux perçus par le robot le ou les signaux d'intérêt [2]. Une fois les signaux séparés, il est alors possible d'en extraire l'information. L'exploitation la plus typique consiste ici à extraire des signaux de paroles pour en déterminer l'origine (*reconnaissance* de locuteur) ainsi que le contenu (*reconnaissance* de parole). Ainsi, toute architecture audio moderne se doit d'être équipée au minimum de ces trois capacités de localisation, extraction et reconnaissance, en témoigne les solutions proposées dans les projets européens récents EARS (KELLERMANN, 2016) et TWO!EARS (RAAKE, 2016) impliquant des robots.

Cependant, le contexte robotique n'apporte pas que des contraintes. A la différence des approches standard de l'analyse de scène audio, les systèmes robotiques sont capables de bouger dans leur environnement. Cette capacité motrice unique différencie ainsi les approches robotiques des méthodes traditionnelles de part leur incarnation en des systèmes actifs, capables de modifier dynamiquement la configuration spatiale de leur(s) capteur(s) audio. Le potentiel de cette approche *active* de l'audition a été identifié très tôt par la communauté (NAKADAI, LOURENS et al., 2000 ; F. WANG et al., 1997). Depuis, un certain nombre de travaux ont clairement démontré comment le mouvement pouvait être exploité pour améliorer l'analyse de la scène sonore (KNEIP et BAUMANN, 2008 ; Y.-C. LU et COOKE, 2010 ; MARTINSON, APKER et BUGAJSKA, 2011 ; BERNARD et al., 2012 ; ZHONG, SUN et YOST, 2016 ; V. NGUYEN et al., 2017). Cette approche active prend d'ailleurs tout son sens au sein du paradigme binaural qui, s'il est exploité de manière statique, n'apporte que peu d'avantage comparativement aux approches s'appuyant sur des réseaux de microphones. De telles considérations réinterrogent les approches proposées ces 15 dernières années, au point que les communautés traditionnelles de l'audio (Traitement du Signal et Acoustique) se sont en partie réappropriées ces questionnements (cf. sessions spéciales aux conférences ICASSP 2015, 2016 et 2017 (BUSTAMANTE, DANÉS et al., 2016 ; EVERS et al., 2017)) en lien avec la robotique. Ainsi, en comparaison avec l'architecture présentée précédemment, les approches récentes incluent donc l'action comme composante à part entière de l'analyse de la scène sonore. Il en résulte une nouvelle architecture, schématisée à la figure 2.2. Ce schéma met en évidence les 2 chemins traditionnels mêlant action et perception (CHAUMETTE, 1998) :

— la voie montante, qui depuis une plateforme possiblement mobile exploite les conséquences du mouvement pour explorer et analyser la scène sonore. On parle alors *d'audition active* ;

— la voie descendante, qui exploite les informations tirées de la scène sonore pour contrôler le système mobile. On parle alors de *commande référencée capteur* (PORTELLO, BUSTAMANTE et al., 2014 ; MAGASSOUBA, BERTIN et CHAUMETTE, 2015 ; MAGASSOUBA, 2016 ; BUSTAMANTE, DANÉS et al., 2016).

Il est intéressant de remarquer que la figure 2.2 indique que toute brique d'analyse audio peut potentiellement être à l'origine de la génération d'une action. Néanmoins, l'essentiel des travaux récents s'est concentré sur la *localisation active* et la façon dont le changement de position du capteur sonore peut être exploité pour remonter à la localisation du ou des sources sonores présentes dans la scène (T. MAY, N. MA et G. J. BROWN, 2015 ; ZHONG, SUN et YOST, 2016 ; ODO et al., 2017). Les autres analyses peuvent intuitivement bénéficier du mouvement pour améliorer leurs performances : typiquement, une source est d'autant

2. On notera qu'il existe néanmoins des approches conjointes de localisation et extraction de source, cf. (DELEFORGE et HORAUD, 2012) par exemple.

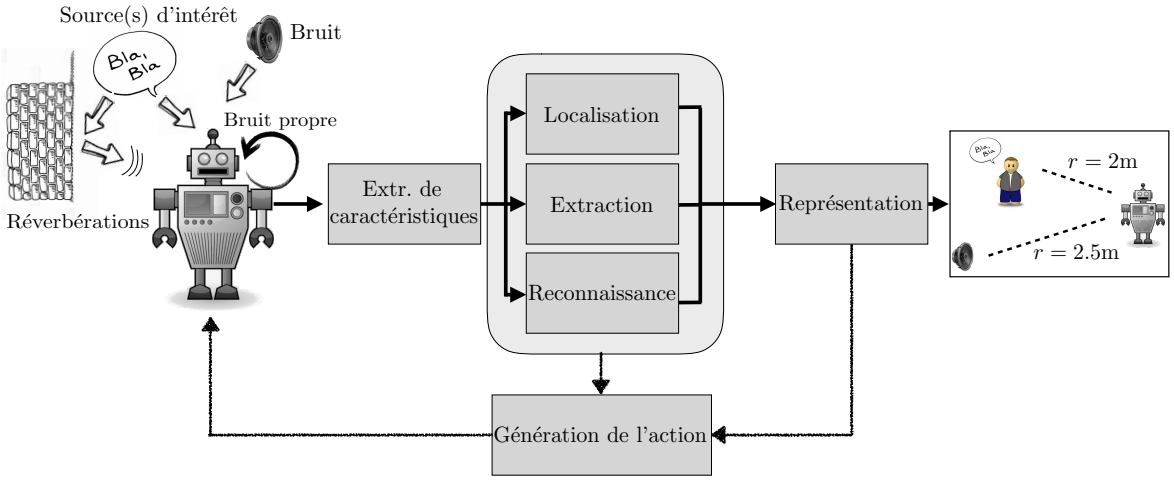


FIGURE 2.2 – Approche *active* de l'audition. A la partie *bottom-up* précédente s'ajoute un retour moteur dont la conséquence est la modification des signaux audio perçus en fonction du mouvement.

mieux reconnue que le système d'analyse en est proche. Pour autant, peu d'auteurs ont encore exploité cette possibilité (KUMON et al., 2010). Enfin, l'élaboration de cartes (audio) de l'environnement peut difficilement s'envisager sans exploration du capteur sonore (SU et al., 2016), à la façon du SLAM exploitant habituellement des données lasers et/ou visuelles. Pour terminer cette rapide contextualisation, il est important de noter que très peu d'approches proposent d'aborder le schéma bouclé de la figure 2.2 dans son ensemble, en traitant à la fois des aspects actifs et de contrôle. On citera dans ce domaine les travaux récents portés par E. Vincent (VINCENT, SINI et CHARPILLET, 2015 ; V. NGUYEN et al., 2017) (contrôle optimal pour la localisation de sources et planification de l'action) et P. Danès (BUSTAMANTE, DANÈS et al., 2017) (contrôle basé information pour la localisation).

Publication

Ces éléments de contextualisation sont en partie abordés dans l'article (ARGENTIERI, DANÈS et SOUÈRES, 2015) publié au sein de la revue "Computer, Speech and Language". Cet article propose un état de l'art détaillé des méthodes de localisation de sources sonores dans un contexte robotique.

2.1.2 Positionnement et ligne de recherche

En 2003, au tout début de nos travaux sur l'audition en robotique (au LAAS-CNRS, Toulouse), les toutes premières contributions cherchaient à reproduire –dans un cadre binaural exclusivement– nos facultés de localisation et reconnaissance. Très rapidement, les limites de ce cadre binaural ont été identifiées, en particulier en terme de robustesse aux différentes contraintes listées précédemment. Les approches à base de réseaux de microphone ont alors été évaluées, adaptées, remaniées de façon satisfaisante, et sont aujourd'hui de loin les plus utilisées. Il est d'ailleurs clair que *qui souhaite un système auditif artificiel performant doit encore aujourd'hui nécessairement se tourner vers des réseaux de microphones.* Nous avons ainsi contribué à l'évolution des systèmes à base de formation de voie pour la localisation de sources sonores, en cherchant à satisfaire la contrainte d'embarquabilité (ARGENTIERI, DANES et SOUÈRES, 2006) (synthèse d'une formation de voie adaptée aux antennes de petite taille) et temporelle (ARGENTIERI et DANÈS, 2007) (approches mixant formation de voie et algorithme MUSIC réduisant drastiquement les temps de calcul). Nous avons également conçu

en parallèle un système hardware à même de déporter un certain nombre d'opérations dédiées à l'analyse de la scène sonore au sein d'un composant FPGA dédié (BONNAL et al., 2010).

Toutes ces approches, bien que prenant en compte explicitement le contexte robotique, n'exploitaient néanmoins pas le mouvement de la plateforme mobile. Ainsi, les opérations de localisation étaient effectuées de manière statique, ou du moins supposées instantanées et sans consolidation temporelle ou prise en compte du contexte. Par ailleurs, les approches humanoïdes –devenues importantes dans les applications envisageant une interaction "naturelle" avec l'homme– ont réinterrogé nos travaux autour des années 2010. Nous nous sommes alors tournés dès 2008 (au sein de l'ISIR, Paris) vers le paradigme binaural, encore très peu exploité en robotique.

C'est dans ce contexte que nous avons monté le projet BINAAHR (BINaural Active Audition for Humanoid Robots, ANR blanche internationale franco-japonaise, 2009-2013) en collaboration avec P. DANÈS (LAAS-CNRS, porteur du projet), et mêlant les équipes françaises de l'ISIR et du LAAS-CNRS avec les équipes japonaises du Prof. OKUNO (Université de Kyoto), du Prof. NAKADAI (Institut de technologies de Tokyo) et Prof. KUMON (Université de Kumamoto). Le projet BINAAHR avait pour objectif d'évaluer des techniques de localisation, de séparation et reconnaissance de sources sonores dans un contexte binaural incluant l'action, ou plus naïvement, ses conséquences (i.e. un changement de position dans une pièce échoïque, par exemple). Ce projet était ainsi notre première tentative d'inclure l'action et ses conséquences au sein d'un système d'analyse audio binaural. Dans ce cadre, seront synthétisés dans la suite de ce manuscrit nos contributions sur les thèmes suivants :

— **localisation de sources sonore** (thèse de Karim Youssef (YOUSSEF, 2013)) :
  - étude et caractérisation des indices binauraux en contexte audio réaliste ;
  - méthode d'apprentissage pour la localisation de sources sonores ;

Ainsi, les travaux présentés dans la suite sur ces sujets s'inscrivent pleinement dans l'architecture traditionnelle montante schématisée à la figure 2.1.

Dans les travaux de Karim YOUSSEF, seul le changement de position du récepteur binaural induit par l'action est pris en compte. C'est clairement une des limites fortes de ses travaux, et à ce titre il est délicat de parler encore de perception active. C'était précisément l'objectif de la thèse d'Alban PORTELLO (PORTELLO, 2013) que d'apporter une formalisation précise de ce problème dans un cadre binaural. Encadrée en grande partie à Toulouse par Patrick DANÈS au cours du projet BINAAHR, cette thèse sera très rapidement abordée au sein de la section 2.3.

Dans ces 2 thèses soutenues en 2013, aucune problématique de commande (au sens "automatique"), même bas-niveau, n'a été abordée. Pour autant, l'enjeu de la chaîne de retour, décidant de l'action à mener pour conduire l'analyse de la scène sonore, est une problématique primordiale. A défaut d'aborder cette question sous l'angle de la commande, nous avons néanmoins cherché à exploiter le mouvement pour guider l'exploration d'un robot dans un environnement inconnu. Ces travaux ont été effectués à l'occasion du projet Européen TWO!EARS [3], qui avait pour objectif d'investiguer l'apport des rétroactions de type "top-down" au sein d'une architecture audio binaurale complète, incluant toutes les étapes d'analyses figurant sur la figure 2.2. Notre contribution à ce projet a ainsi porté sur :

— **interprétation active de la scène sonore** : définition d'un étage haut-niveau d'analyse de la scène sonore, fusionnant les informations audio et visuelles disponibles sur un robot mobile à l'aide de mouvements de tête pouvant être modulés selon la *congruence* des stimulis (thèse de B. Cohen-Lhyver (COHEN-LHYVER, 2017)).

Une partie de nos contributions à ces deux problématiques (localisation et interprétation) est synthétisée dans la suite de ce manuscrit.

3. 2013–2016, projet FET open.

2.2 Contributions à la localisation binaurale de source sonore

Cette section décrit nos contributions en localisation de sources sonores dans un contexte binaural et robotique. Comme mentionné précédemment, un défit important de cette problématique est la robustesse aux conditions acoustiques, qui seront systématiquement supposées réalistes, i.e. incluant du bruit et des réverbérations. De plus, nous avons proposé d'inclure le mouvement et ses conséquences (changement de position du récepteur binaural dans une pièce échoïque et modification en conséquence de la perception audio de la scène) de façon à caractériser, modéliser et exploiter ses effets dans le processus de localisation. Cette section s'articule comme suit. Dans un premier temps, nous proposons un bref rappel sur la localisation des sons, en particulier chez l'homme. Sur cette base, nous étudierons tout d'abord les conséquences des réverbérations sur les indices acoustiques extraits des signaux binauraux. Il s'agit ici d'identifier parmi l'ensemble des techniques d'estimations de ces indices, laquelle est la plus robuste (au sens défini dans la suite) aux réverbérations. Alors, nous proposons dans une seconde contribution d'utiliser les méthodes identifiées préalablement pour localiser, via des méthodes d'apprentissage, un son dans un environnement depuis un robot mobile placé en différents endroits au sein d'une pièce réverbérante. L'exploitation explicite du mouvement dans le processus de localisation sera mentionnée dans la section 2.3, où des méthodes de filtrage stochastique permettront l'estimation de la position d'une source au cours du mouvement.

2.2.1 Comment localiser un son dans un contexte binaural ?

Les mécanismes régissant la localisation des sons chez l'homme ont fait l'objet de beaucoup d'études (et continuent encore à être objet de recherches aujourd'hui (AFGHAH et al., 2017 ; BEDNAR, BOLAND et LALOR, 2017)). Néanmoins ses principes de base sont maintenant largement connus depuis plus d'un siècle et les études de Lord Rayleigh (RAYLEIGH, 1907), complétées depuis par de larges considérations psychoacoustiques (MIDDLEBROOKS et M. GREEN, 1991). L'ensemble de ces connaissances constitue naturellement la base des développements sur l'audition binaurale en robotique. Ainsi, les différentes étapes successives nécessaires à la localisation d'un son peuvent être résumées en les étapes suivantes :

— **Propagation :** le son se propage depuis la source sonore à localiser jusqu'à atteindre le récepteur binaural. Sur son chemin, l'onde sonore va interagir avec le corps, la tête, et l'oreille externe du robot. Les effets de cette interaction peuvent être capturés par la fonction de transfert de la tête (Head Related Transfer Function, ou HRTF), pouvant inclure dans un contexte robotique les effets combinés des différents éléments constitutifs du "corps" du robot. On suppose donc ici que l'environnement n'a aucun effet sur la propagation : la HRTF ne caractérise alors que l'effet physique du robot, supposé seul dans un environnement anéchoïque (i.e. dans un des conditions acoustique proches du champ libre) ;

— **Caractéristiques audio :** ensuite, les sons captés par le récepteur binaural sont analysés pour en extraire des caractéristiques. Chez l'homme, ces *indices monauraux* (extraits depuis seulement un des 2 signaux disponibles) ou *binauraux* (extraits depuis les 2 signaux binauraux) ont fait l'objet de nombreuses études, et sont également exploités en robotique comme caractéristiques bas niveau pour la localisation ;

— **Algorithmie :** enfin, sur la base des caractéristiques binaurales précédentes, on peut envisager n'importe quelles méthodologies (apprentissage, analytiques, estimation, etc.) pour permettre de mettre en relation ces indices audio avec leur origine spatiale.

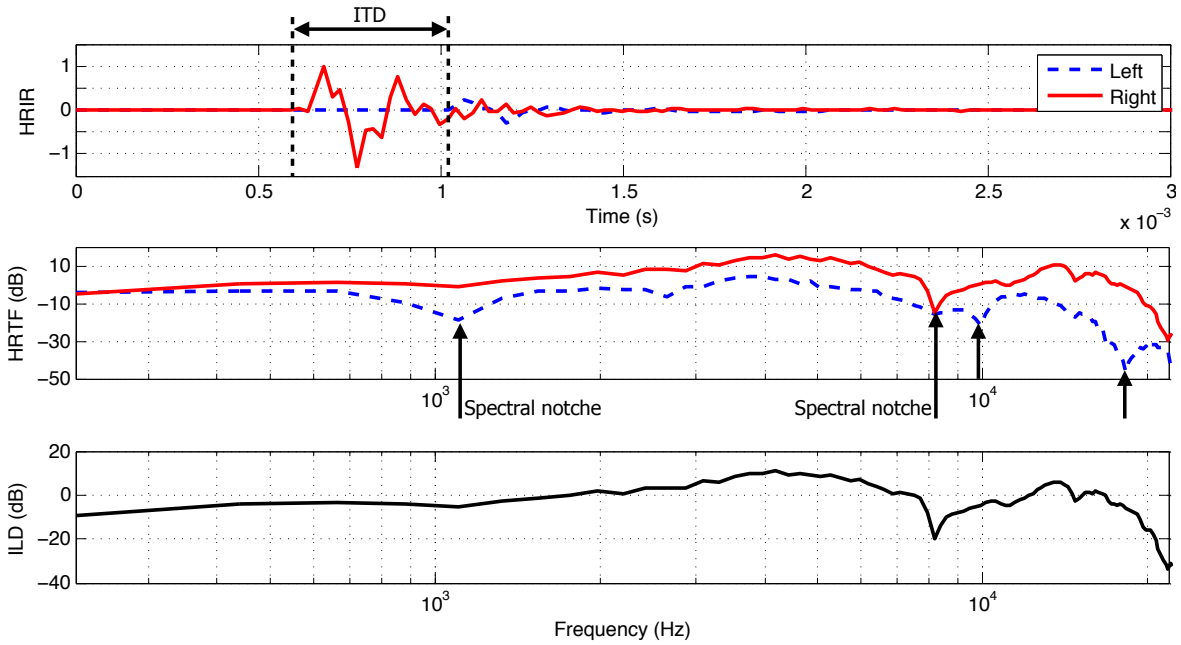


FIGURE 2.3 – Représentation des effets de la tête sur les signaux binauraux, capturés en fonction du temps (HRIR, en haut) ou de la fréquence (HRTF, au milieu). La différence d'amplitude entre les signaux gauche et droite (ILD, en bas) constitue un des indices binauraux les plus utilisés. Figure tirée de (ARGENTIERI, DANÈS et SOUÈRES, 2015).

2.2.1.1 Notations

Dans toute la suite, les deux signaux binauraux (notés comme provenant de la gauche et de la droite de la tête par exemple) sont obtenus par échantillonnage à la fréquence f_e et seront notés $l[n]$ et $r[n]$ respectivement, où n désigne l'indice temporel. Leurs pendants fréquentiels seront respectivement notés $L[k]$ et $R[k]$, avec k l'indice fréquentiel. En pratique, ces deux représentations fréquentielles sont classiquement obtenues à l'aide d'une transformée de Fourier discrète sur des fenêtres de N points, de sorte que la fréquence $f[k] = kf_e/N$. Ces deux signaux binauraux ont pour origine une source sonore s placée en les coordonnées polaires $(r_s, \theta_s, \varphi_s)$, où le centre de la tête est pris par convention comme origine du repère. Ainsi, r_s représente la distance à la source, θ_s l'azimut de la source (dans le plan horizontal) et φ_s son élévation (dans le plan vertical). Cette source s émet un signal $s_s[n]$, de contenu fréquentiel $S_s[k]$. On supposera enfin que la position $(\theta_s, \varphi_s) = (\pi/2, 0)$ correspond à une source placée devant la tête.

2.2.1.2 Fonction de transfert de la tête et indices audio pour la localisation

La HRTF met en relation le signal émis par la source sonore et les deux signaux binauraux, de sorte que

$$\begin{cases} L[k] = H_l(f[k], r_s, \theta_s, \varphi_s)S_s[k] \\ R[k] = H_r(f[k], r_s, \theta_s, \varphi_s)S_s[k] \end{cases}, \tag{2.1}$$

où H_l et H_r désignent respectivement les HRTF gauche et droite, représentées pour une position de source sonore latérale droite sur la figure 2.3. On peut y constater que d'une part, l'amplitude de la HRTF droite est plus élevée que celle de gauche, traduisant ainsi l'existence d'une *différence interaurale d'amplitude* (ILD). La source placée à droite produit en effet un son d'amplitude plus faible sur l'oreille gauche, celle-ci n'ayant pas de vue directe sur la source. Cette différence d'amplitude n'est pas constante et dépend de la fréquence, comme

le stipule les lois de diffraction acoustique à la surface d'un solide. De la même façon, on peut mettre en évidence sur les contreparties temporelles des HRTFs, appelées *réponses impulsionnelles de la tête* (HRIR), l'existence d'un délais entre elles. Il s'agit ici d'une *différence interaurale en temps* (ITD), causée par la différence de chemin à parcourir entre la source sonore et les deux oreilles. A nouveau, cette différence temporelle n'est pas en général constante et dépend, pour les mêmes raisons que l'ILD, de la fréquence. Enfin, on peut remarquer la présence de zéros (*spectral notches*) plus ou moins marqués dans les HRTFs gauche et/ou droite. L'ensemble de ces caractéristiques constitue la base des indices utilisés chez l'homme pour localiser un son : les différences interaurales sont majoritairement influencées par la position azimutal de la source sonore, tandis que la position des zéros en fréquence sur chacune des 2 oreilles (on parle alors d'indice *monaural*) est dépendante de son élévation. Enfin, l'influence de la distance de la source sur tous ces indices est faible au delà d'environ 1m, de sorte qu'il est particulièrement difficile d'en estimer la valeur de manière statique. D'autres caractéristiques peuvent néanmoins être extraites des signaux binauraux, sous l'hypothèse de l'existence de réverbérations qui habituellement dégradent fortement les performances de l'analyse de scène. L'importance de ces réverbérations peut être en partie évaluée grâce au *temps de réverbération* RT60 qui mesure le temps mis par le signal mesuré pour diminuer de 60dB une fois la source dans l'environnement éteinte. Le RT60 dépend principalement de la taille de la pièce et de l'absorption acoustique des murs (KINSLER et al., 1999). Plus sa valeur est importante, et plus l'environnement sera qualifié de réverbérant. Sur la base de cette propriété acoustique, l'estimation de la distance peut se baser sur le rapport de l'énergie directe sur réverbérante (DRR), ce rapport venant à diminuer d'environ 25dB lorsque la distance à l'émetteur double (ZAHORIK, 2002). L'ensemble de ces indices, naturellement exploités en robotique, est défini et caractérisé dans la suite.

2.2.2 Définition et caractérisation des indices binauraux

Tandis qu'il y a un consensus évident sur l'origine et la signification des différents indices binauraux ou monauraux mentionnés précédemment, nous avons fait le constat à l'occasion des débuts de la thèse de Karim YOUSSEF qu'il n'en était pas de même sur la façon de les estimer. Plus précisément, des définitions différentes en ont été données dans de nombreux travaux. Ce constat, fait au début de ses travaux en 2010, a également été fait par la communauté Acoustique (KATZ et NOISTERNIG, 2014). Il nous semblait dès lors important de tenter de synthétiser l'ensemble des définitions proposées par les différentes Communautés, mais également de mettre en évidence leurs propriétés au sein d'une étude statistique comparative. Ces deux points sont abordés dans les 2 sous-parties suivantes.

2.2.2.1 Définitions et propriétés

Dans toute la suite, nous nous restreindrons à l'étude des caractéristiques exploitées pour estimer la position d'une source sonore dans le plan horizontal via son azimut θ_s et sa distance r_s. L'élévation de la source sera donc supposée nulle ou, à défaut, connue d'avance. Les méthodes listées dans la suite s'appuient pour la plupart sur un fenêtrage des signaux sur N points, de sorte que les indices binauraux sont estimés toutes les N/f_e secondes (si aucun recouvrement temporel n'est effectué lors de l'analyse). Par commodité, toute référence à ce fenêtrage est ôté des notations.

Indices binauraux pour l'estimation de l'azimut Travailler sur la localisation horizontale revient à estimer un retard (l'ITD) et une différence d'amplitude (l'ILD) entre les deux

signaux binauraux gauche et droite $l[n]$ et $r[n]$. Dès lors, il parait naturel d'exploiter la corrélation croisée pour l'estimation de ce retard, ainsi que le rapport des énergies de ces deux signaux, de sorte qu'on définisse respectivement ITD_{CC} et ILD par

$$\begin{cases} ITD_{\text{CC}} = \dfrac{1}{f_e} \arg\max_m C_{lr}[m], \text{ avec } C_{lr}[m] = \displaystyle\sum_{n=0}^{N-m-1} l[n+m]r[n], \\ ILD = 10\log_{10} \dfrac{\sum_{n=0}^{N-1} l[n]^2}{\sum_{n=0}^{N-1} r[n]^2}. \end{cases} \tag{2.2}$$

Ces deux définitions sont probablement les plus utilisées dans la littérature. Elles souffrent néanmoins de plusieurs limites : la précision du calcul de l'ITD est faible et limitée à la valeur d'une période d'échantillonnage, et les résultats d'ITD comme d'ILD ne dépendent pas de la fréquence. Le premier point peut être résolu en exploitant des techniques d'interpolation et/ou de suréchantillonnage (H. D. KIM et al., 2008 ; LIU et Y. WANG, 2010). Le second point peut être partiellement amélioré en utilisant la *corrélation croisée généralisée*, de sorte que

$$ITD_{\text{GCC}} = \frac{1}{f_e} \arg\max_m GCC_{lr}[m], \text{ avec } GCC_{lr}[m] = IFFT(G[k]L[k]R^*[k]), \tag{2.3}$$

où $G[k]$ désigne une pondération fréquentielle permettant de donner plus ou moins d'importance à certaines composantes spectrales des deux signaux binauraux. De nombreux choix sont possibles ici, parmi lesquelles la pondération PhaT (pour Phase Transform) $G_{\text{PhaT}}[k]$ définie par

$$G_{\text{PhaT}}[k] = \frac{1}{|L[k]||R[k]|}, \tag{2.4}$$

ou encore les pondérations Roth (ROTH, 1971), SCoT (CARTER, NUTTALL et CABLE, 1973) ou HT (HANNAN et THOMSON, 1973). Le résultat de cette analyse, bien que nécessitant un passage dans le monde fréquentiel, produit néanmoins toujours un unique scalaire, et le résultat obtenu reste indépendant de la fréquence. De toutes les pondérations possible, PhaT est probablement la plus exploitée dans la littérature et permet, via le blanchiment des signaux, de gommer toute différence d'amplitude préalablement au calcul de corrélation. Si ce blanchiment permet ainsi de ne considérer que les phases des signaux perçus, il est aussi à l'origine d'un manque évident de robustesse au bruit dans l'environnement.

La prise en compte explicite de la dépendance spectrale des signaux binauraux passe par leur décomposition en fréquence préalablement à toute analyse. Une première approche évidente consiste ainsi à déterminer les transformées de Fourier discrètes des deux signaux, et à déterminer ainsi les indices binauraux $ILD_{\text{FFT}}[k]$ et $IPD_{\text{FFT}}[k]$, où l'IPD désigne la Différence Interaurale en Phase, pendant fréquentiel de l'ITD. On a alors

$$\begin{cases} IPD_{\text{FFT}}[k] = \arg(L[k]) - \arg(R[k]), \\ ILD_{\text{FFT}}[k] = 20\log_{10} \dfrac{|L[k]|}{|R[k]|}. \end{cases} \tag{2.5}$$

De part le nombre de points utilisés pour fenêtrer les signaux, ces deux indices s'avèrent très fortement redondants. Il s'agit alors d'en réduire la dimension, via 2 approches naïves néanmoins souvent exploitées. Il est possible de moyenner les indices obtenus par bandes de fréquences successives (approche "FFT_1"), ou de moyenner les spectres sur ces mêmes

bandes préalablement au calcul des indices binauraux (approche "FFT_2"). On a alors

$$\begin{cases} IPD^{(i)}_{\text{FFT}_1} = \sum_{k=k_l[i]}^{k_u[i]} IPD_{\text{FFT}}[k], \\ ILD^{(i)}_{\text{FFT}_1} = \sum_{k=k_l[i]}^{k_u[i]} ILD_{\text{FFT}}[k], \end{cases} \tag{2.6}$$

où l'indice $i \in [0 \dots L-1]$ représente la i$^{\text{ième}}$ bande de fréquences comprise entre $k_l[i]$ et $k_u[i]$, avec $k_l[i] < k_u[i]$, et $k_u[i] < k_l[i+1]$. On a également

$$\begin{cases} IPD^{(i)}_{\text{FFT}_2} = \arg\left(\sum_{k=k_l[i]}^{k_u[i]} L[k]\right) - \arg\left(\sum_{k=k_l[i]}^{k_u[i]} R[k]\right), \\ ILD^{(i)}_{\text{FFT}_2} = 20\log_{10}\dfrac{|\sum_{k=k_l[i]}^{k_u[i]} L[k]|}{|\sum_{k=k_l[i]}^{k_u[i]} R[k]|}. \end{cases} \tag{2.7}$$

Il est également possible d'envisager une décomposition en fréquence plus proche de celle effectuée au sein de l'oreille interne de l'homme, à l'aide de filtres dit *cochléaires* reproduisant la décomposition spectrale opérée au sein de la cochlée. Il s'agit ici d'exploiter des bancs de filtres dont les caractéristiques bande-passante/fréquence centrale s'inspirent du profil vibratoire de la membrane basilaire à l'origine du déplacement des cellules ciliées, transducteurs mécano-électrique de l'audition. Pour ce faire, il est courant d'utiliser des filtres gammatone, de réponse impulsionnelle $h(t)$ donnée par

$$h(t) = at^{n-1}e^{-2\pi Bt}\cos(2\pi f_0 t + \phi), \tag{2.8}$$

où n désigne l'ordre du filtre, f_0 sa fréquence centrale et B sa bande passante. Ainsi, si L filtres cochléaires droite et gauche sont utilisés, on dispose de $2L$ signaux de sortie $l^{(i)}[n]$ et $r^{(i)}[n]$ qui permettent de déterminer les indices binauraux selon

$$\begin{cases} ITD^{(i)}_{\text{AFE}} = \arg\max_m C^{(i)}_{lr}[m], \text{ avec } C^{(i)}_{lr}[m] = \sum_{n=0}^{N-m-1} \text{AFE}(l^{(i)}[n+m])\text{AFE}(r^{(i)}[n]), \\ ILD^{(i)}_{\text{AFE}} = 10\log_{10}\dfrac{\sum_{n=0}^{N-1}\text{AFE}(l^{(i)}[n])^2}{\sum_{n=0}^{N-1}\text{AFE}(r^{(i)}[n])^2}, \end{cases} \tag{2.9}$$

où la fonction AFE(.) (pour *Auditory Front End*) désigne un ensemble de traitements supplémentaires possiblement appliqués aux signaux de sortie des filtres cochléaires. Il s'agit alors, en plus de modéliser l'effet de la cochlée via sa décomposition fréquentielle, d'inclure des modèles de transduction neuronale. Nous proposons d'évaluer dans la suite :

— l'AFE proposé par (T. MAY, S. VAN DE PAR et A. KOHLRAUSCH, 2011), consistant en une rectification demi-onde $\Pi(.)$, suivie d'une compression en racine-carrée, i.e. $\text{AFE}_1(.) = \sqrt{\Pi(.)}$,

— un AFE inspiré de (BERNSTEIN et TRAHIOTIS, 1996 ; FALLER et MERIMAA, 2004) et appliquant une compression d'enveloppe, une rectification demi-onde, mise au carré et filtrage passe-base $W[.]$ d'ordre 4 à une fréquence de coupure de 425Hz, i.e. $\text{AFE}_2(.) = W[\Pi(H(.)^{-0.8})^2]$ avec $H(.)$ l'enveloppe du signal obtenue en déterminant le module du signal analytique,

— et un AFE sans aucun traitement autre que le filtrage cochléaire, de sorte que $\text{AFE}_3(.) = Id(.)$.

Indices pour l'estimation de la distance L'estimation de la distance reste un problème difficile. Longtemps, cette problématique a été abordée en cherchant à exploiter les mêmes indices binauraux que ceux utilisés pour la localisation en azimut (typiquement, ITD/IPD et ILD), alors que leur sensibilité à la distance est faible. A la place, l'estimation du rapport d'énergie directe sur réverbérante (DRR) a permis des avancées notables. Une première façon simple d'en estimer la valeur est l'algorithme d'égalisation-annulation proposé dans (Y. C. LU et COOKE, 2010). Prenant arbitrairement comme référence le signal gauche de sortie $l^{(i)}[n]$ de la $i^{\text{ème}}$ bande de fréquence d'intérêt, on peut égaliser le signal droite $r^{(i)}[n]$ en compensant sa différence d'amplitude et son décalage temporel, créant ainsi le signal égalisé $r_{\text{eq}}^{(i)}[n]$ défini par

$$r_{\text{eq}}^{(i)}[n] = \sqrt{\frac{\sum_{n=0}^{N-1} l^{(i)}[n]^2}{\sum_{n=0}^{N-1} r^{(i)}[n]^2}} r^{(i)}\left[n - \text{ITD}_{\text{AFE}_3}^{(i)} f_e\right]. \tag{2.10}$$

Évidemment, l'égalisation n'est pas parfaite. Sous l'hypothèse que cette erreur d'égalisation est totalement liée à la présence des réverbérations, on peut alors estimer l'énergie réverbérée $R^{(i)}$ dans la $i^{\text{ème}}$ bande de fréquence via

$$R^{(i)} = \sum_{n=0}^{N-1} |l^{(i)}[n] - r_{\text{eq}}^{(i)}[n]|^2. \tag{2.11}$$

Enfin, comme l'énergie $S^{(i)}$ de la $i^{\text{ème}}$ bande de fréquence est $S^{(i)} = \sum_{n=0}^{N-1} l^{(i)}[n]^2$ et vérifie par hypothèse $S^{(i)} = D^{(i)} + R^{(i)}$, où $D^{(i)}$ représente l'énergie directe de la $i^{\text{ème}}$ bande de fréquence, on a alors

$$DRR_{\text{EQ}}^{(i)} = \frac{S^{(i)} - R^{(i)}}{R^{(i)}}. \tag{2.12}$$

Cette reformulation de l'estimation du DRR selon la fréquence permet ainsi d'obtenir un indice binaural susceptible de capturer les réflexions et absorptions acoustiques à l'origine des réverbérations, toutes deux étant des fonctions de la fréquence et des matériaux équipant la salle.

Une autre façon de tenir compte des propriétés de la pièce où est situé le robot est d'exploiter des modèles de corrélation spatiale acoustique, utilisés à l'origine dans des contexte d'antennerie. Nous avons proposé, dans le cadre de la thèse de K. YOUSSEF, d'étudier le modèle proposé par (HIOKA et al., 2011) qui, sous certaines hypothèses (ondes planes uniquement, HRTF négligée, composante réverbérante diffuse et faiblement corrélée à la composante directe, principalement), permet d'écrire les auto et inter corrélations des signaux binauraux $R_{ll}[k]$, $R_{lr}[k]$, $R_{rl}[k]$ et $R_{rr}[k]$, avec $R_{lr}[k] = \mathbb{E}\left[L[k]R^*[k]\right]$, sous la forme

$$\begin{aligned} &\bar{\mathbf{R}}[k] = (R_{ll}[k], R_{lr}[k], R_{rl}[k], R_{rr}[k])^T = \mathbf{F}[k]\mathbf{P}[k], \\ &\text{avec } \mathbf{P}[k] = \begin{pmatrix} P_D[k] \\ P_R[k] \end{pmatrix} \text{ et } \mathbf{F}[k] = \begin{pmatrix} 1 & d_{lr} & d_{lr} & 1 \\ 1 & r_{lr} & r_{lr} & 1 \end{pmatrix}^T, \end{aligned} \tag{2.13}$$

où $P_D[k]$ et $P_R[k]$ représentent respectivement les densités spectrales de puissance des composantes directes et réverbérantes des signaux gauche et droite, $d_{lr} = 1/d_{rl} = e^{j2\pi k \frac{f_e \text{ITD}}{Nc}}$ et $r_{lr} = r_{rl} = \text{sinc}\left(2\pi k \frac{a f_e}{Nc}\right)$, avec a la distance entre les 2 oreilles et c la vitesse de propagation du son. Ainsi, la matrice $\mathbf{P}$ peut être estimée via $\widetilde{\mathbf{P}}[k] = (\tilde{P}_D[k], \tilde{P}_R[k])^T = \mathbf{F}^+[k]\bar{\mathbf{R}}[k]$, où $^+$

représente la pseudo-inverse de Moore-Penrose, de sorte qu'on obtienne

$$\mathrm{DRR}_{\mathrm{SCM}}^{(i)} = 10\log_{10}\left(\frac{\sum_{k=k_l[i]}^{k_u[i]}\tilde{P}_D[k]}{\sum_{k=k_l[i]}^{k_u[i]}\tilde{P}_R[k]}\right). \tag{2.14}$$

Ainsi, sur la base d'une estimée de la matrice de corrélation $\bar{\mathbf{R}}[k]$, il devient possible d'estimer via l'équation (2.14) une estimée du DRR.

Enfin, en plus du DRR, la cohérence des signaux binauraux $\gamma[k]$–qui mesure la corrélation linéaire entre les deux signaux gauche et droite en fonction de la fréquence– est un indice supplémentaire renseignant sur la distance à la source, déjà utilisé dans (VESA, 2009). Il est défini selon

$$\gamma[k] = \frac{|G_{lr}[k]|^2}{G_{ll}[k]G_{rr}[k]}, \tag{2.15}$$

avec $G_{lr}[k] = \langle L^*[k]R[k]\rangle$, $G_{ll}[k] = \langle |L[k]|^2\rangle$ et $G_{rr}[k] = \langle |R[k]|^2\rangle$, où $\langle Q[k]\rangle = 0.5Q[k]_{t-1} + 0.5Q[k]_{t-1}$ avec t l'indice de la trame omis dans les notations.

2.2.2.2 Étude comparative

Ayant listé les différentes méthodes permettant d'estimer les indices binauraux, il s'agit maintenant de les comparer. Lorsqu'on cherche à comparer différentes caractéristiques, il est souvent d'usage de les comparer via leur utilisation concrète par un algorithme commun. Typiquement, nous pourrions imaginer comparer ces différentes caractéristiques en utilisant un algorithme de localisation donné (basé modèles acoustiques, apprentissage, etc.) et en comparant les performances de localisation obtenues. Nous proposons ici de définir à la place une métrique statistique simple permettant de quantifier la dispersion des différents indices en fonction de l'importance des réverbérations, évaluée pour rappel via le RT60. Ainsi, l'indice le moins sensible aux réverbérations, i.e. dont la dispersion variera le moins lorsque le temps de réverbération augmente, sera vraisemblablement le plus robuste quant aux conditions acoustiques.

Métrique statistique Dans toute la suite, nous proposons de séparer les indices acoustiques en L groupes positionnels, chaque groupe l étant défini par son azimut ou sa distance. En notant m_l le nombre de trames temporelles prises en compte dans l'estimation des matrices suivantes, et $M_L = \sum_{l=1}^{L} m_l$ le nombre total de trames pour tous les groupes, on peut définir :

- la matrice moyenne de dispersion intragroupe $\mathbf{W} = \frac{1}{M_L}\sum_{l=1}^{L} m_l\mathbf{W}_l$, avec $\mathbf{W}_l$ la matrice de covariance du $l^{\text{ième}}$ groupe positionnel ;
- la matrice moyenne de dispersion intergroupe $\mathbf{B} = \frac{1}{M_L}\sum_{l=1}^{L} m_l(\boldsymbol{\mu}_l - \boldsymbol{\mu})^T(\boldsymbol{\mu}_l - \boldsymbol{\mu})$, où $\boldsymbol{\mu}_l$ et $\boldsymbol{\mu}$ représentent respectivement le centre du $l^{\text{ième}}$ groupe positionnel et le centre de la totalité des groupes ;
- le Lambda de Wilks () $\wedge$, qui mesure la séparation des centres des groupes, avec

$$\wedge = \frac{\det \mathbf{W}}{\det\left(\mathbf{B} + \mathbf{W}\right)}. \tag{2.16}$$

Le Lambda de Wilks est un scalaire compris entre 0 et 1, dont la valeur va diminuer lorsque la dispersion intragroupe diminue relativement à la dispersion totale intra et intergroupe. Ainsi, à de faibles valeur de $\wedge$ correspond un "conditionnement" des indices, relativement à leur dispersion fonction de la réverbération, de meilleur "qualité".

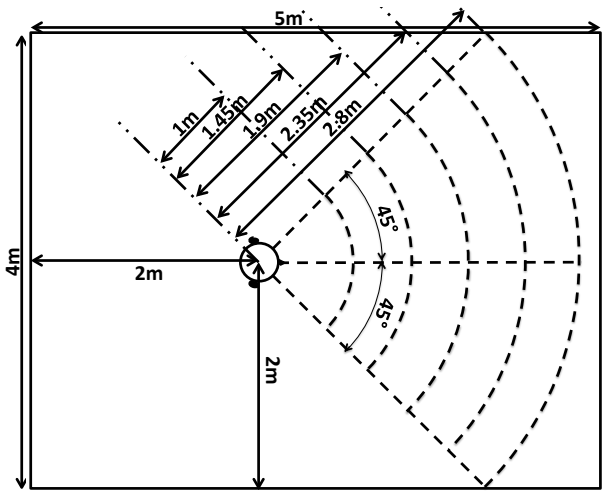
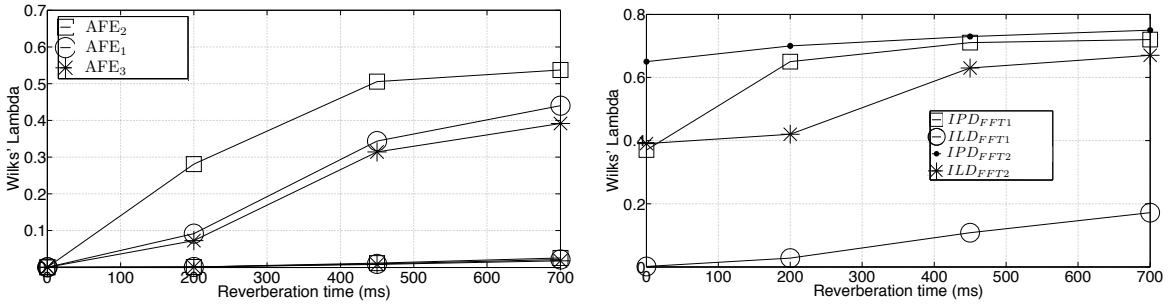


FIGURE 2.4 – Conditions expérimentales pour constituer la base de données d'évaluation des indices binauraux. La source, placée à une élévation nulle, occupe des positions comprises entre $-45°$ et $+45°$ par pas de $5°$, à des distances comprises entre 1m et 2.8m par pas de 45cm. La base de donnée est donc constituée de 19 valeurs angulaires et 5 valeurs de distance. Avec les matériaux simulés, la pièce possède un RT_{60} d'environ 200ms à 1kHz. Figure tirée de (YOUSSEF, 2013).

Base de données Afin d'évaluer le Lambda de Wilks pour chacun des indices définis précédemment, et pour différentes conditions réverbérantes, nous avons construit une base de données à l'aide du logiciel Roomsim (CAMPBELL, PALOMÄKI et BROWN, 2005) simulant l'acoustique d'une pièce rectangulaire dont les paramètres (taille, matériau, absorption, taux d'humidité, etc.) peuvent être précisément réglés. Sur cette base, un récepteur binaural (un simulateur de tête et torse KEMAR) est placé à une position donnée dans la pièce. Le simulateur exploite alors la méthode des images (ALLEN et BERKLEY, 1979) et la base de données des HRTF du mannequin binaural KEMAR pour produire les deux signaux binauraux desquels seront extraits l'ensemble des caractéristiques binaurales listées précédemment. Les conditions précises simulées sont synthétisées à la figure 2.4. Les coefficients d'absorption des murs sont ensuite modifiés de façon à modifier les conditions acoustiques, produisant un temps de réverbération RT_{60} à 1kHz de 450ms et 700ms, en plus des conditions nominales (RT_{60}=200ms) et anéchoïques. La source sonore émet un signal de parole issu d'une base de données, fenêtré via des fenêtres rectangulaires de 1024 points (durée = 23ms pour $f_e = 44.1$kHz), et passé préalablement à toute analyse dans un détecteur d'activité vocale élémentaire basée sur l'énergie du signal.

Résultats La figure 2.5a trace la valeur de $\wedge$ pour les 3 modèles d'AFE listés en §2.2.2.1. Pour rappel, AFE_1 et AFE_2 représentent deux modèles de complexité croissante, tandis que AFE_3 représente les données sans traitement préalable à l'extraction des indices selon l'équation (2.9). Dans cette évaluation, 30 filtres cochléaires sont utilisés. Le premier commentaire qu'appelle cette figure concerne la grande différence de comportement des indices temporels (ITD) et d'amplitude (ILD). Ces derniers possèdent en effet des valeurs associées de $\wedge$ bien plus faibles que pour leurs pendants temporels, témoin de leur plus grande robustesse aux réverbérations. Travailler avec l'ITD en environnement réverbérant apparaît donc comme une mauvais idée, ce qui semble cohérent avec l'idée que la cohérence temporelle entre les deux signaux binauraux se dégrade d'autant plus que le temps de réverbération est important. De la même façon, on constate que travailler avec le modèle le plus simple, i.e. ne reproduisant aucun effet de transduction neuronale, concoure à améliorer là aussi la discrimination spatiale des indices binauraux. La compression d'information induite par



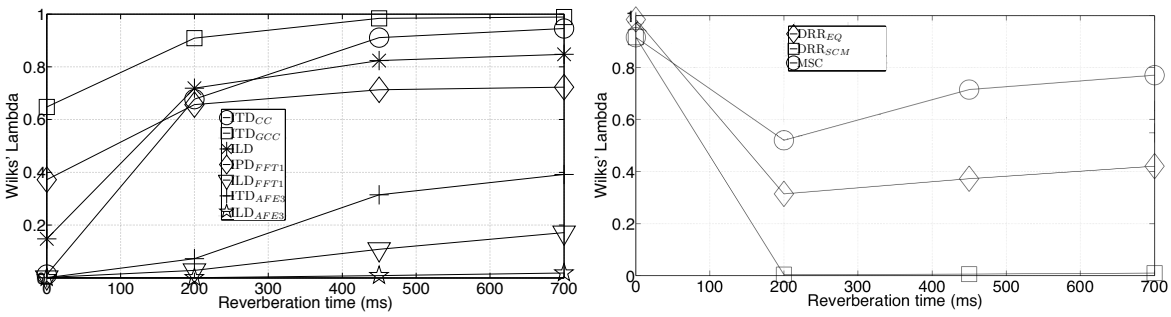
(A) AFE (B) FFT

FIGURE 2.5 – Comparaison des valeurs de $\wedge$ pour *(gauche)* différents modèles auditifs *(droite)* différents moyennages des FFT, produisant chacun des estimées des ITDs (courbes du dessus) et ILDs (courbes du dessous). Figure tirée de (YOUSSEF, 2013).

ces différents modèles est probablement à l'origine de ce comportement. Nous avons également comparé les 2 stratégies de moyennage de FFT proposées aux équations (2.6) et (2.7) au sein de la figure 2.5b. Par analogie aux bancs de filtres précédents, 30 bandes de fréquences successives sont également utilisées. Il apparaît clairement que les valeurs de $\wedge$ associées à la stratégie FFT_1 sont beaucoup plus faibles que dans la seconde stratégie : travailler avec la moyenne des indices binauraux plutôt que sur les indices obtenus par moyennage des spectres conduit à une plus faible dispersion dans l'espace des données en fonction du temps de réverbération. C'est donc cette stratégie qui sera retenue pour comparaison dans la suite.

L'ensemble des indices listés en §2.2.2.1 est comparé au sein de la figure 2.6a. Il y apparaît à nouveau que les indices en amplitude présentent généralement une plus grande robustesse aux réverbérations se traduisant par une valeur associée de $\wedge$ plus faible. Par ailleurs, l'ensemble des méthodes sans décomposition spectrale produit également de moins bons résultats (i.e. un $\wedge$ plus haut) que les approches impliquant une décomposition fréquentielle explicite. Et des deux décompositions fréquentielles étudiées, celle s'appuyant sur les filtres gammatone est à l'origine d'une moins grande dispersion des indices en fonction de la réverbération. L'allure des réponses en fréquence des filtres gammatone, non régulièrement espacés en fréquence et à bande passante variable, explique vraisemblablement ce meilleur comportement.

Enfin, les 3 indices binauraux sensibles à la distance sont comparés au sein de la figure 2.6b. Comme précédemment, les indices sont calculés sur 30 bandes de fréquences successives. A la différence des indices binauraux utilisés pour la localisation en azimut, il



(A) Indices binauraux pour l'azimut (B) Indices binauraux pour la distance

FIGURE 2.6 – Comparaison des valeurs de $\wedge$ pour *(gauche)* différents modèles auditifs *(droite)* différents moyennages des FFT, produisant chacun des estimées des ITDs (courbes du dessus) et ILDs (courbes du dessous).Figure tirée de (YOUSSEF, 2013).

apparaît que les performances discriminatives en environnement anéchoïques des indices de distance sont particulièrement mauvaises ($\wedge \approx 1$), ce qui est tout à fait logique à la vue de leur définition (le champ réverbéré étant quasi nul). Il est intéressant de constater que pour autant, les réverbérations dégradent également $\wedge$ lorsqu'elles augmentent. Au final, l'approche basée sur la corrélation spatiale (donnant lieu calcul du DRR_{SCM}) semble la plus robuste. Modéliser la diffusion du champ acoustique, même sur la base d'hypothèses fortes pas nécessairement vérifiées dans les simulations proposées, permet d'aboutir à un indice binaural particulièrement intéressant pour l'estimation de la distance en comparaison des autres approches.

Publication

La contribution synthétisée dans cette sous-section a donné lieu à l'article (K. YOUSSEF, S. ARGENTIERI et J. L. ZARADER, 2012), publié et présenté au sein des sessions spéciales "Robot Audition" de la conférence IEEE/RSJ IROS.

2.2.3 Apprentissage de la localisation

L'étude précédente a permis d'illustrer la grande variété de définitions et estimateurs des mêmes caractéristiques de localisation du signal binaural. Sur cette base, nous avons proposé un critère permettant de sélectionner la "meilleure" approche relativement à sa sensibilité aux conditions expérimentales. La suite logique de ce travail de Karim YOUSSEF a consisté à exploiter ces indices, mais avec toujours l'idée de confronter les méthodologies de localisation à des contextes réalistes. Dans ce cadre, et alors que les approches par apprentissage pour de l'analyse audio étaient encore peu répandues en Robotique, la thèse de Karim YOUSSEF a été l'occasion de nous focaliser sur un algorithme simple d'estimation de l'azimut et de la distance de la source s'appuyant sur un réseau de neurones. Synthétisés dans la suite, ces travaux mettent en évidence les conséquences 1/d'un placement différent du récepteur binaural au sein de l'environnement (par exemple suite à une action du robot), et 2/des réverbérations sur l'estimation de la position de la source.

2.2.3.1 Paramétrisation du problème

Nous proposons dans la suite d'exploiter l'architecture proposée à la figure 2.7. Elle met en évidence les 2 étapes successives traditionnelles de ce type d'approche, i.e. une extraction de caractéristique suivie d'un réseau de neurones. Ces deux parties sont détaillées dans la suite.

Extraction de caractéristiques : sur la base de l'étude statistique précédente, nous choisissons d'extraire les caractéristiques de localisation en sortie de filtres gammatones, i.e. nous exploitons la stratégie AFE_3 définie à l'équation (2.9) fournissant $ITD^{(i)}_{\text{AFE}_3}$ et $ILD^{(i)}_{\text{AFE}_3}$ sur L bandes de fréquences. Pour l'estimation de la distance, nous nous appuyons sur le $\text{DRR}^{(i)}_{\text{SCM}}$ défini à l'équation (2.14) déterminé sur la base d'une analyse en fréquence sur L bandes. Ainsi, les deux vecteurs de caractéristiques utilisés respectivement pour l'estimation de l'azimut et distance sont donnés par

$$\begin{cases} V^{\text{az}}[t] = [ITD^{(1)}_{\text{AFE}_3}, \ldots, ITD^{(15)}_{\text{AFE}_3}, ILD^{(16)}_{\text{AFE}_3}, \ldots, ILD^{(30)}_{\text{AFE}_3}], \\ V^{\text{dist}}[t] = [\text{DRR}^{(1)}_{\text{SCM}}, \ldots, \text{DRR}^{(30)}_{\text{SCM}}]. \end{cases} \tag{2.17}$$

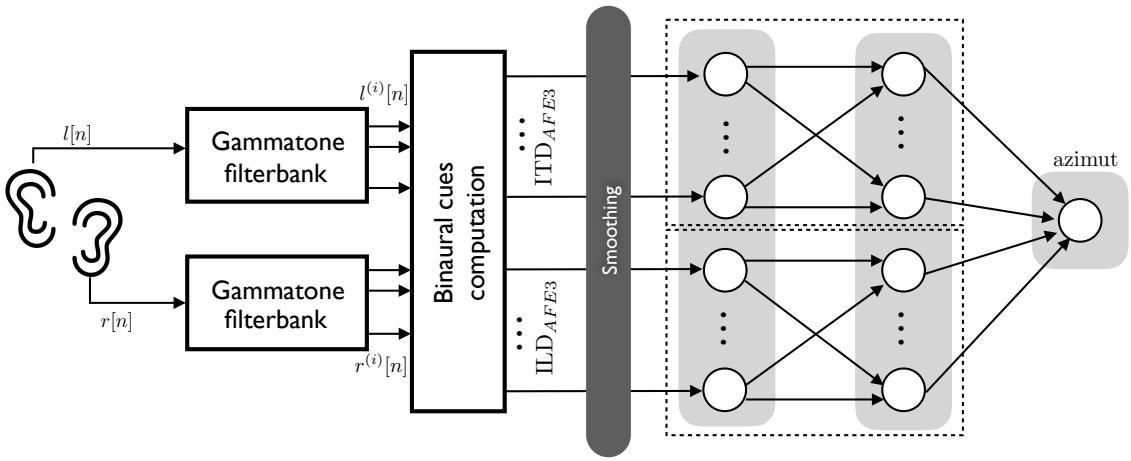


FIGURE 2.7 – Architecture du système de localisation proposé (pour l'azimut). N filtres cochléaires sont utilisés pour déterminer les indices ITD, ILD et DRR par bande de fréquence. Sur cette base, un réseau de neurones partiellement connecté exploite l'ITD et l'ILD pour fournir une estimée de l'azimut, tandis qu'un second réseau (totalement connecté) fournit lui une estimée de la distance.

L'équation (2.17) met en évidence deux aspects. D'une part, $L = 30$ bancs de filtres ont été sélectionnés. L'influence du choix de ce nombre de banc de filtre a été détaillé dans (YOUSSEF, 2013), et la valeur de 30 y figure comme le meilleur compromis entre la dimension des deux vecteurs de caractéristiques et le gain de performances évalué au sens de la métrique (2.16). D'autre part, les deux vecteurs font apparaître une dépendance temporelle via l'indice de fenêtre t. Ceci est le résultat d'un moyennage temporel effectué sur les vecteurs de caractéristiques afin d'améliorer la robustesse aux réverbérations. Le dimensionnement de ce moyennage a également été étudié de sorte que 10 vecteurs de caractéristiques successifs sont utilisés pour obtenir le vecteur moyen utilisés dans la suite. Cette valeur est le meilleur compromis entre une dynamique temporelle suffisante pour garantir une bonne réactivité du système et ses performances statistiques. Pour terminer, on constate que le vecteur de caractéristiques V^{az} est constitué d'éléments hétérogènes : des ITDs exprimés en s associés aux fréquences les plus faibles, et des ILDs exprimés en dB associés aux fréquences les plus hautes. Ceci est le résultats des propriétés connues de ces 2 indices : l'ITD fournit des valeurs non ambiguës pour les faibles fréquences, tandis que l'ILD n'est significatif que pour les plus hautes.

Réseau de neurones : deux réseaux de neurones sont exploités de manière indépendante pour estimer l'azimut et la distance de la source. Le premier réseau, dédié à l'estimation angulaire, est un réseau à connexions partielles. La nature du vecteur d'entrée, mêlant ITD et ILD, justifie ce choix. Le réseau utilisé pour l'estimation de distance est quant à lui entièrement connecté. Après étude comparative des propriétés de ces réseaux, tous deux sont dotés d'une couche cachée de 14 neurones, pour 30 neurones d'entrée.

Génération des données simulées : comme précédemment, nous exploitons un environnement de simulation réaliste pour générer les signaux binauraux desquels sont extraits les vecteurs de caractéristique. Dans notre contexte robotique, il nous faut générer des données représentatives des différents scenarii envisageables : le robot peut bouger dans l'environnement (de multiples positions doivent être testées), changer d'environnement (différentes conditions acoustiques doivent être envisagées), et chercher à localiser une source émettant depuis n'importe quelle position (influence croisée entre les estimations angulaire et de distance). Ainsi, nous proposons d'exploiter le même type de base de données générées dans la

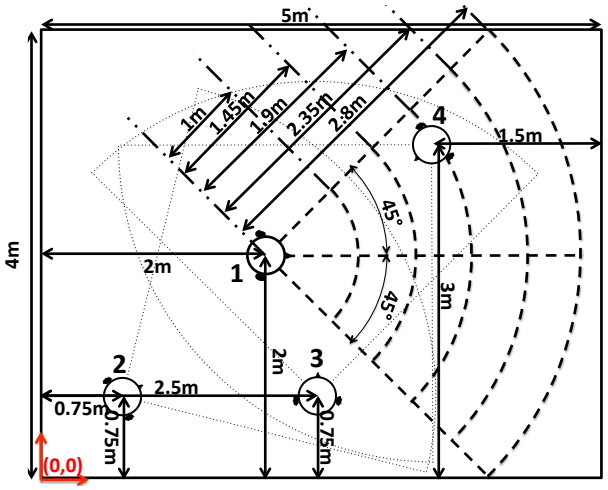


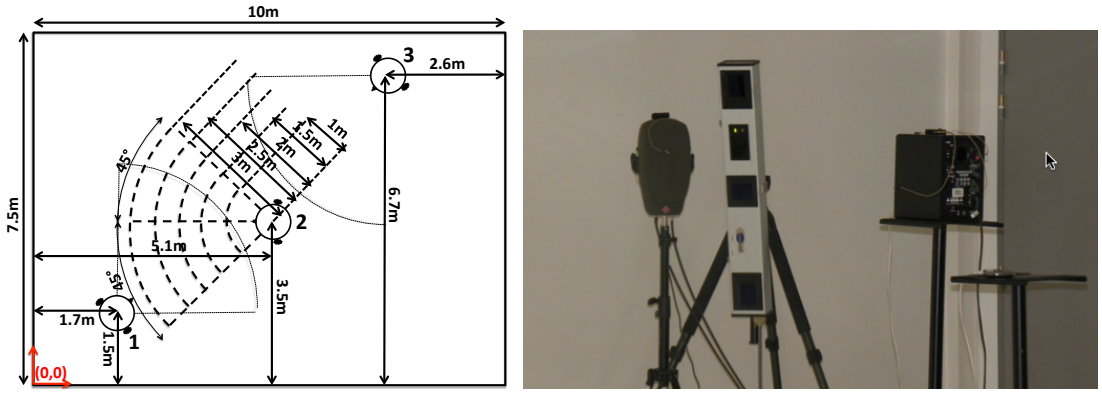
FIGURE 2.8 – Conditions expérimentales pour constituer la base de données d'évaluation de l'algorithme de localisation. La source, placée à une élévation nulle, occupe des positions comprises entre -45° et $+45^\circ$ par pas de 5°, à des distances comprises entre 1m et 2.8m par pas de 45cm, du récepteur binaural. La tête binaurale est placée en 4 positions et orientations différentes, et les matériaux utilisés dans la simulation sont ajustés pour modifier le temps de réverbération. Figure tirée de (YOUSSEF, 2013).

sous-section §2.2.2.2, dont les paramètres sont synthétisés sur la figure 2.8. La source simulée émet soit un bruit blanc, soit des signaux de parole. Les données issues de ces simulations sont ensuite réparties en 3 parties : une base d'apprentissage (environ 3000 exemples), une base de validation croisée (environ 1000 exemples) pour éviter le sur-apprentissage, et une base de validation (environ 5000 exemples). On notera, compte tenu des capacités de simulation disponibles, la faible dimension des bases de données constituées. Ce choix se justifie dans la mesure où il est rare de disposer, expérimentalement, d'un nombre important de mesures effectuées dans les mêmes conditions (plusieurs environnements acoustiques, plusieurs positions du récepteur audio, etc.) Il s'agit donc aussi d'évaluer dans quelle mesure l'approche proposée permet d'envisager, dans un contexte où il est difficile d'acquérir expérimentalement beaucoup de données, un algorithme d'apprentissage pour la localisation de source. Ce point est tout particulièrement critique pour l'établissement de la base de données réelle suivante.

Génération des données réelles : des signaux binauraux ont également été enregistrés à l'occasion de la constitution d'une base de données expérimentale, enregistrée depuis un tête binaurale KU100 de Neumann au sein d'une pièce de $10 \times 7.5 \times 2.8$m non échoïque, voir figure 2.9. Un haut-parleur, jouant le rôle de source sonore, émet les mêmes sons que ceux utilisés en simulation. De façon à disposer de la réalité terrain, un système de capture de mouvements est utilisé pour mesurer avec une grande précision les positions relatives source/récepteur. 3 positions du récepteur binaural sont utilisées, et pour chacune des positions 95 enregistrement binauraux sont effectués, chacun correspondant à une position/distance relative de la source au récepteur.

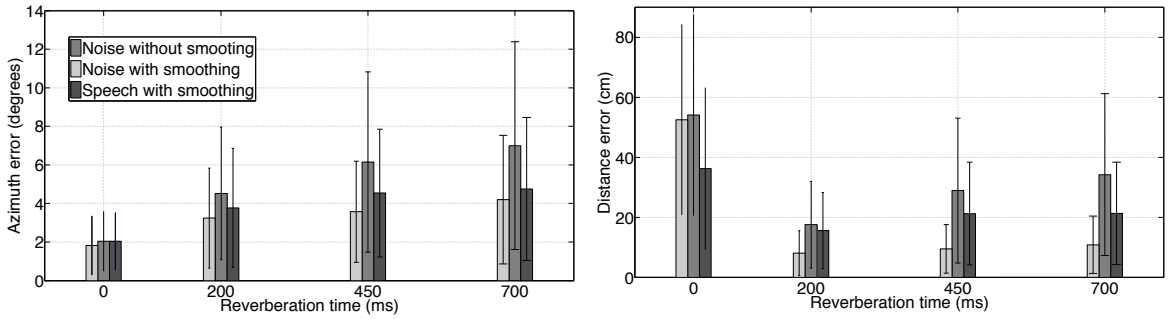
2.2.3.2 Résultats

Nous proposons ici une synthèse des évaluations effectuées en 2 points : une étude "meilleur cas", effectuée dans des conditions d'apprentissage et de test identiques, et une étude des capacités de généralisation du réseau.



(A) Conditions expérimentales (B) Photo du dispositif

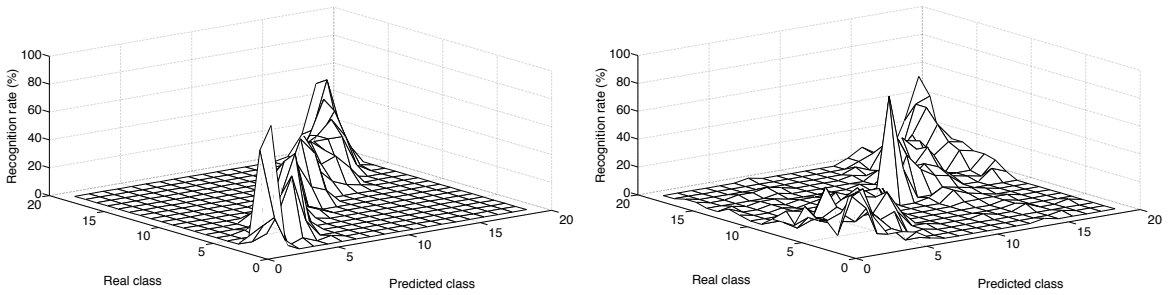
FIGURE 2.9 – Conditions expérimentales pour l'établissement de la base de donnée réelles. Un haut-parleur, placée à une élévation nulle, occupe des positions angulaires comprises entre $-45°$ et $+45°$ par pas de $5°$ (précision de $0.18°$), à des distances comprises entre 1m et 3m par pas de 50cm (précision de 1.46cm), de la tête binaurale KU100. Ce récepteur est placé en 3 positions et orientations différentes au sein d'une pièce échoïque. Figure tirée de (YOUSSEF, 2013).



(A) Estimation de la distance (B) Estimation de l'azimut

FIGURE 2.10 – Erreur d'estimation de la distance (gauche) et de l'azimut (droite) pour différents temps de réverbérations. Figure tirée de (YOUSSEF, 2013).

Évaluation pour des conditions d'apprentissage et de test identiques Dans cette partie, les données simulées sont utilisées. Dans l'objectif d'évaluer les capacités d'estimation des paramètres spatiaux du système proposé dans un cas idéal, les conditions acoustiques (spécifiées via la valeur de temps de réverbération RT60), et la position du récepteur binaural, sont identiques entre les phases d'apprentissage et de test. Les résultats obtenus sont alors ceux présentés sur la figure 2.10. On peut y voir que l'erreur d'estimation de l'azimut est inférieure ou égale à $2°$ dans des conditions anéchoïque, et ce quel que soit le type de son émis par la source. Les performances se dégradent comme attendu lorsque le RT60 augmente, cette augmentation pouvant être contenue grâce à l'introduction du moyennage temporel des caractéristiques mentionné précédemment. L'estimation de la distance est quant à elle très mauvaise en environnement anéchoïque de part l'exploitation, au sein de la définition même du DRR, du champ réverbéré inexistant dans un tel environnement. Là encore, plus le RT60 augmente, moins précise est l'estimation de distance obtenue. Néanmoins, l'erreur sur la distance reste inférieur à 20cm dans les conditions les plus réverbérantes. Au final, dans ces conditions optimales, la méthode proposée permet de localiser avec une marge d'erreur faible une source sonore dans un environnement réverbérant. A noter que nous avons également conduit une étude comparative entre notre approche et celle proposée au même moment dans (T. MAY, S. VAN DE PAR et A. KOHLRAUSCH, 2011 ; T. MAY, S. VAN DE PAR et A. KOHLRAUSCH, 2012), qui propose d'exploiter des mixtures de gaussiennes (GMM)



(A) Matrice de confusion (réseau de neurones) (B) Matrice de confusion (GMM)

FIGURE 2.11 – Matrices de confusion pour les 2 approches réseau de neurones/GMM, obtenues dans les mêmes conditions : base de données (signaux de paroles) identique, RT60=450ms. Figure tirée de (YOUSSEF, 2013).

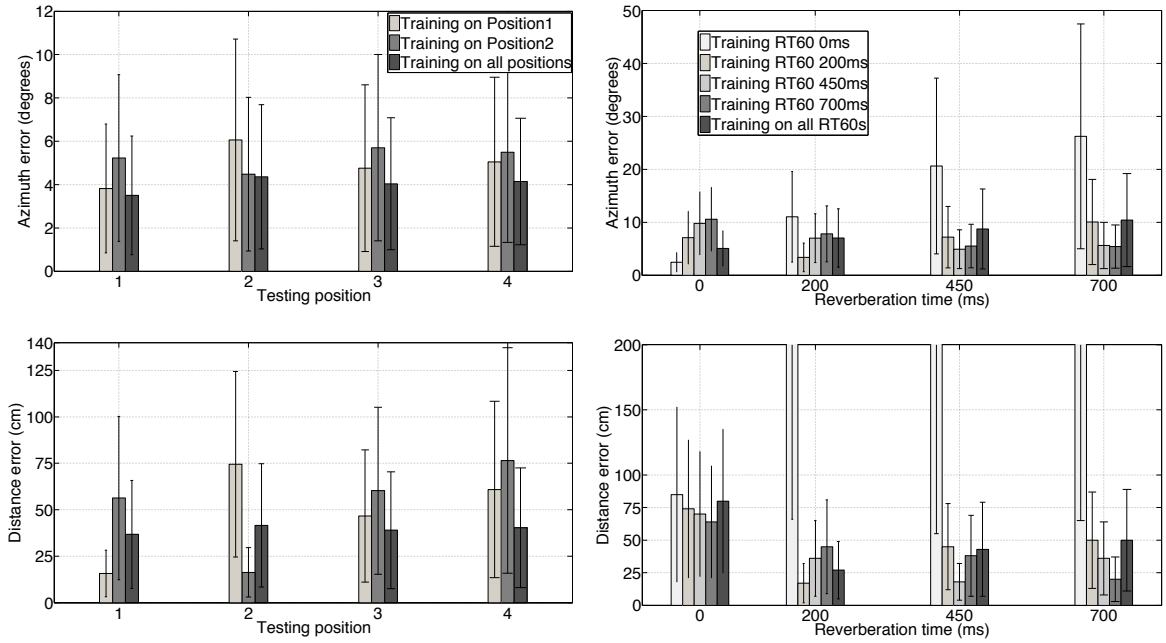
en lieu et place du réseau de neurone. Afin d'en comparer les sorties (azimut en degré vs. appartenance à une classe), les sorties du réseau de neurones sont regroupées en classes positionnelles de $5°$ (par exemple, si l'azimut estimé appartient à l'intervalle $[-2.5°, 2.5°]$, il sera associé à la classe "$0°$"). Comme dans l'article original, les GMM sont paramétrées de telles sortes qu'une mixture de 15 gaussiennes est associée à chaque sortie de filtre gammatone, pour 19 classes positionnelles de $5°$ comprises entre $-45°$ et $45°$. Les 2 approches sont exploitées dans exactement les mêmes conditions (apprentissage, test, bases de données, signaux). En résumé, l'étude comparative montre :

— dans les conditions mentionnées précédemment, le système à base de GMM produit de meilleurs taux de reconnaissances (au sens des classes positionnelles) dans des conditions faiblement réverbérantes. Néanmoins, la matrice de confusion obtenue pour les GMM montre que les erreurs de classification se répartissent très largement autour de la vraie position de la source. Au contraire, la matrice de confusion obtenue pour le réseau de neurone montre un comportement beaucoup plus homogène, voir la figure 2.11 ;

— ce premier point est très probablement lié au fait que la base de données exploitée est volontairement constituée d'un nombre d'exemples "limité", réaliste au regard des exploitations expérimentales robotiques envisagées. Dès lors, augmenter la taille de la base de données permet aux GMM d'obtenir de bien meilleurs résultats ... néanmoins inatteignables expérimentalement. Une autre solution consiste également à réduire la complexité du système à base de GMM (réduction de nombre de gaussiennes par mixture et nombre de banc de filtres), au détriment toujours des performances.

Capacités de généralisation Sur la base du réseau de neurones évalué précédemment, plusieurs études portant sur la capacité de généralisation de l'approche ont été menées. En effet, dans un cadre robotique où la plateforme binaurale est susceptible de se déplacer, les positions depuis lesquelles sont captées les données binaurales influencent grandement les performances de localisation. De la même façon, si les conditions de réverbération évoluent (typiquement si le robot change de pièce), il faut nous assurer de la robustesse de l'approche à ce type de changement.

Pour un RT60 fixe de 200ms, l'erreur d'estimation de l'azimut et de la distance en fonction des différentes positions occupées par le récepteur binaural lors de la phase d'apprentissage est représentée sur la figure 2.12a. On peut y constater que non seulement changer de position par rapport à la phase d'apprentissage dégrade effectivement les performances d'estimation en distance et azimut, mais également que certaines positions (ici la



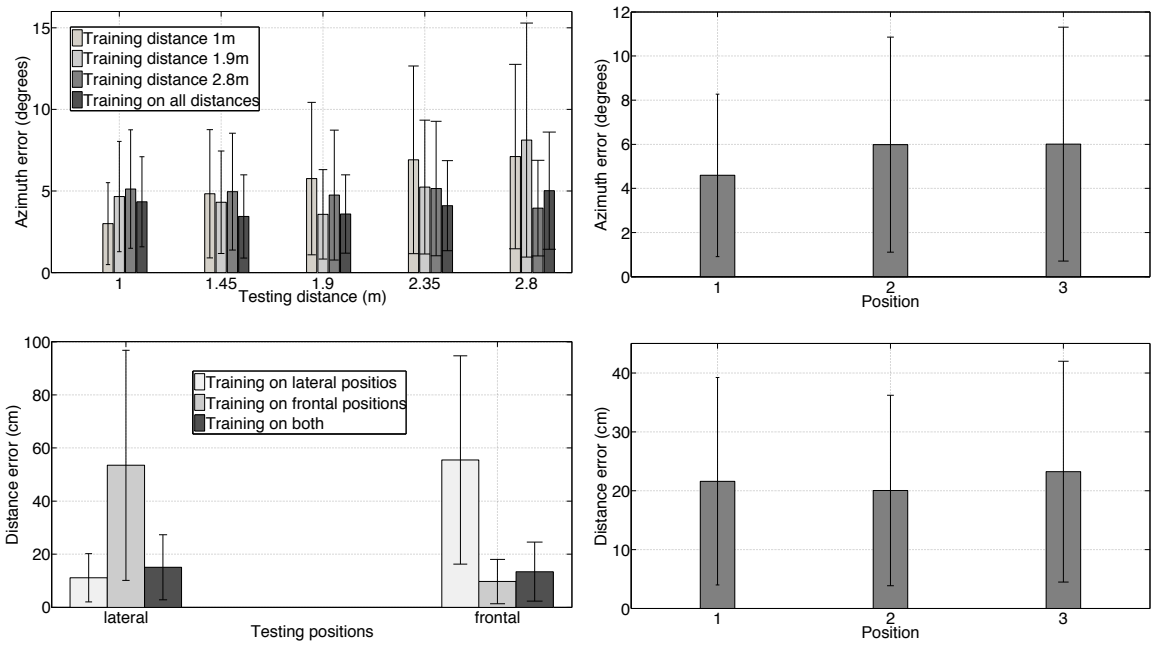
(A) Influence de la position d'apprentissage (B) Influence du RT60 lors de l'apprentissage

FIGURE 2.12 – Étude de l'influence de la position (gauche) et des conditions de réverbération (droite) sur l'erreur d'estimation de l'azimut (haut) et de la distance (bas). Figures tirées de (YOUSSEF, 2013).

position n°2) conduisent à des capacités de généralisation moindres. Typiquement, la position 2 semble en effet la plus difficile d'un point de vue acoustique (récepteur dans un coin de la pièce). En moyenne, les performances sont meilleures lors d'un apprentissage multi-positionnel, comme on pouvait s'y attendre. Mais cela implique donc la création d'une base de données pour de multiples positions potentielles du récepteur audio.

Pour le capteur binaural placé en position 1 pour les phases d'apprentissage et test, différentes conditions acoustiques sont maintenant présentées au système. Dans un premier temps, l'apprentissage est effectué pour une valeur de RT60, et le test est réalisé pour une autre valeur. Les résultats montrent là encore la nécessité de réaliser un apprentissage multi-conditionnel, qui propose alors les meilleurs capacités de généralisation. A noter que comme on pouvait s'y attendre, apprendre dans des conditions anéchoïque (RT60=0s) conduit aux capacités de généralisation les plus faibles, en particulier pour la distance.

Enfin, la distance à la source influence possiblement la qualité de l'estimation de l'azimut (et vice-versa). Cela est d'autant plus vrai que les indices interauraux classiques sont parfois utilisés comme caractéristiques permettant l'estimation de la distance, cf. §2.2.2.1. Il s'agit par exemple d'effectuer l'apprentissage du réseau à une certaine distance du récepteur binaural, et d'évaluer la qualité de l'estimation angulaire du système pour d'autres distances. Cette étude est résumée sur la figure 2.13a. On peut y voir une certaine sensibilité de l'estimation de l'azimut à la distance utilisée lors de la phase d'apprentissage d'autant plus importante que cette distance est faible. Les effets de champ proche peuvent expliquer ce comportement (du moins pour une distance de 1m). Là encore, les meilleurs résultats sont obtenus après un apprentissage multiconditionnel sur un ensemble de distances proches à lointaines. De la même façon, la figure 2.13a (bas) indique qu'il vaut mieux effectuer l'apprentissage de la distance depuis des azimuts différents, frontaux (azimut compris entre $-20°$ et $+20°$) comme latéraux (azimut compris entre $-20°$ et $-45°$ ou son symétrique).



(A) Influence mutuelle de la distance et de l'azimut (B) Résultats expérimentaux

FIGURE 2.13 – Étude de l'influence mutuelle de la distance et de l'azimut sur les capacités de généralisation de l'approche (gauche).Résultats expérimentaux (droite). Figures tirées de (YOUSSEF, 2013).

Résultats expérimentaux Pour terminer cette synthèse des résultats obtenus, la base de données expérimentales présentée à la figure 2.8 est exploitée dans le même type d'études que précédemment. Du fait de sa dimension limitée, seuls les résultats de l'étude positionnelle sont présentés à la figure 2.13b, dans le cas où l'apprentissage et le test ont lieu à la même position. Il y apparaît que le système est capable d'estimer correctement la position de la source avec une erreur moyenne de 4 à 6 degrés en azimut, et environ 20cm en distance. La position 1 montre une meilleure estimation angulaire : l'atténuation apportée par des rideaux à côté réduisent les réverbérations, ce qui en même temps conduit à une moins bonne estimation de la distance. La position 2, au centre de la pièce, permet une meilleure estimée de la distance : le champ diffus est probablement plus proche du modèle exploité dans les indices acoustiques, expliquant ainsi ce meilleur résultat.

Publication

La contribution synthétisée dans cette sous section a donné lieu à l'article (K. YOUSSEF, S. ARGENTIERI et J. L. ZARADER, 2013) publié et présenté au sein des sessions spéciales "Robot Audition" de la conférence IEEE/RSJ IROS.

2.2.3.3 Conclusion

Cette étude exhaustive d'un système très classique à base de réseau de neurone pour l'estimation de la position d'une source sonore met en évidence le besoin de disposer de données exhaustives susceptible de balayer l'ensemble des dimensions du problème. On peut envisager alors deux problématiques :

— comment, sur la base de données synthétiques générées en nombre depuis des simulateurs acoustiques réalistes (reproduisant fidèlement bruit ambiant, réverbération, physique de la salle, etc.) peut-on espérer passer "dans le monde réel" avec de bonnes performances ? Cette question, classique en apprentissage artificiel, reste un sujet de recherche à part entière, et est à notre connaissance pas ou peu abordée dans un contexte audio ;

— comment, sur la base de données expérimentalement identifiées, peut-on généraliser les apprentissages effectués depuis un nombre de données limitées ? Notre participation au projet européen TWO !EARS impliquant des chercheurs reconnus en audition binaurale a montré qu'il est encore nécessaire de collecter, identifier, mesurer énormément de données spécifiques à l'environnement courant du robot pour espérer obtenir, dans ces mêmes conditions, des résultats corrects d'estimation de position. Cette approche permet néanmoins, par exemple après identification des HRIR du robot, d'enrichir les données expérimentales de données simulées de manière plus réaliste que dans des simulateurs physiques.

Enfin, l'approche abordée ici se base sur une approche "quasi-statique" de l'audition. L'action du robot, si elle conduit à un déplacement du récepteur binaural –et donc de ses conditions de captations audio–, n'est pas réellement exploitée à ce stade. C'est justement un des objectifs des travaux de thèse d'Alban PORTELLO et Benjamin COHEN-LYVER que d'exploiter explicitement les capacités d'action du robot. Ces travaux sont présentés dans les section 2.3 et 2.4 respectivement.

2.2.4 Apprentissage multimodal de la localisation de source sonore

Les travaux précédents avaient principalement pour objectif d'évaluer, de quantifier et d'exploiter *uniquement* des informations audio pour procéder à la localisation d'une source sonore dans un environnement réaliste. Pour autant, la plupart des plateformes robotiques modernes sont également équipées d'autres modalités extéroceptives, et en particulier de capacités visuelles. Dès lors, disposer d'informations multimodales interroge l'idée d'utiliser seulement des données audio pour effectuer l'analyse d'une scène sonore. Cette idée, prémisse des travaux de Benjamin COHEN-LHYVER, est évaluée dans la suite.

Il est clair que l'intérêt du contexte robotique pour une tâche d'analyse de scène sonore réside dans la capacité qu'a un robot à traiter des flux sensoriels de natures différentes. Pourtant, très peu de travaux en audition robotique cherchent à exploiter cette capacité, avec l'idée que quand on cherche à localiser une source sonore, on utilise alors uniquement les données audio. Convaincus que du point de vue du robot et de son comportement il n'était pas utile d'exprimer la tâche de localisation en terme angulaires objectifs (azimut, distance, etc.), nous avons redéfini la tâche de localisation audio en terme de co-localisation visuelle : apprendre la position d'une source sonore revient alors à essayer de la localiser dans l'image. Cette opération est représentée sur la figure 2.14. Un haut parleur est repéré dans l'image issue de la caméra équipant le robot par les coordonnées pixels de sa boite englobante. Ce haut parleur émet en parallèle un son qui est capté par le récepteur binaural, et les indices binauraux classiques en sont extraits. Un réseau de neurone apprend alors la correspondance entre ces indices binauraux et la position de la source dans l'image. Le tracé à droite de la figure 2.14 superpose la trajectoire réelle du haut parleur dans l'image (rouge) avec celle estimée par les indices binauraux (bleu). On y constate une certaine imprécision verticale, liée à la nature des indices binauraux utilisés, mais la localisation horizontale est

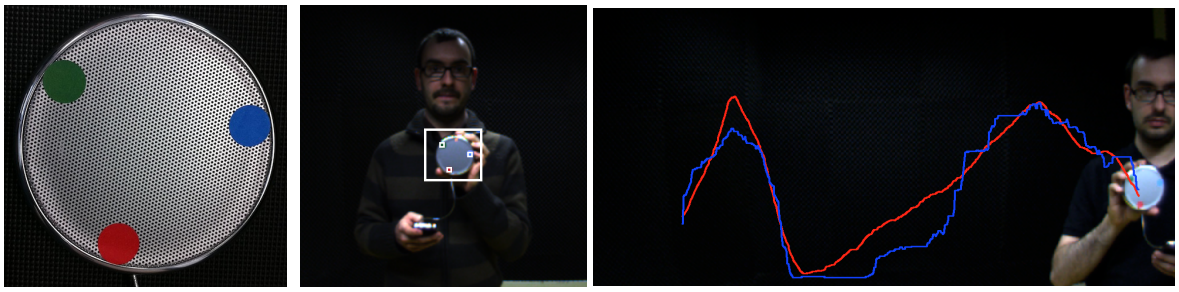


FIGURE 2.14 – Approche multimodale de la localisation : (gauche) haut-parleur, repéré par des marqueur, et (milieu) traqué dans l'image. (Droite) projection de la localisation binaurale audio au sein de l'image : (rouge) trajectoire réelle, (bleu) trajectoire estimée. Figure tirée de (YOUSSEF, ARGENTIERI et ZARADER, 2012).

suffisamment précise pour déterminer avec une erreur de seulement 20 pixels la position horizontale du haut-parleur dans l'image.

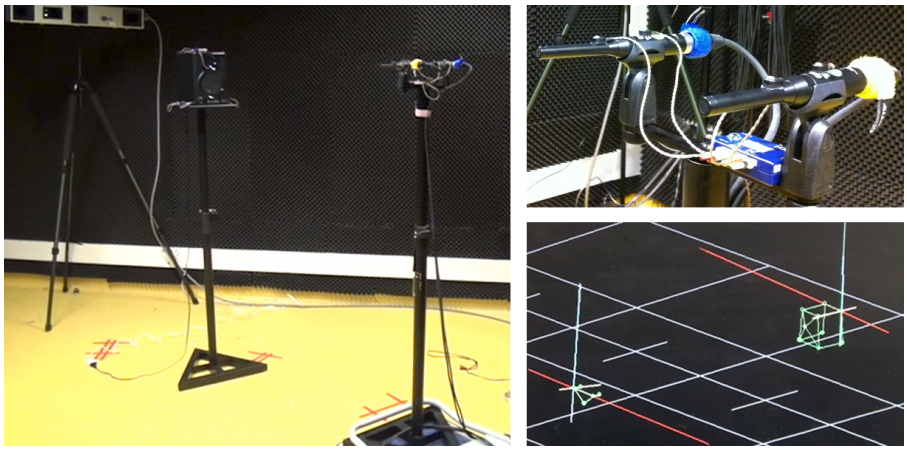
Publication

Cette contribution a donné lieu à l'article (YOUSSEF, ARGENTIERI et ZARADER, 2012) publié et présenté au sein de la conférence IEEE ICASSP.

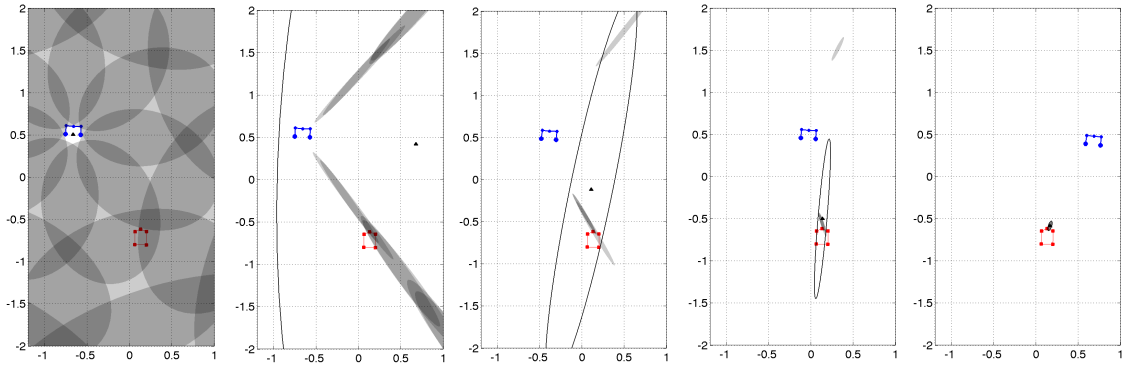
2.3 Localisation binaurale active de sources sonores

En parallèle des travaux menés à l'occasion de la thèse de Karim YOUSSEF, nous avons eu l'occasion de co-encadrer la thèse d'Alban PORTELLO (PORTELLO, 2013) effectuée au sein de l'Université Paul Sabatier à Toulouse et majoritairement suivie par P. Danès au LAAS-CNRS. A ce titre, ses travaux sont très brièvement décris dans la suite. Ils illustrent néanmoins comment inclure de manière plus explicite l'action au sein d'une tâche de localisation de source sonore.

Les travaux d'Alban PORTELLO ont porté sur l'estimation active de la position de sources sonores à l'aide de filtres stochastiques. Il s'agissait alors d'établir une représentation d'état du robot doté de capacités de perception auditive modélisant la façon dont les ordres moteurs modifient la position/orientation de la source sonore par rapport au récepteur binaural. Sur cette base, et en établissant comment ce même mouvement est susceptible de modifier la perception audio, ont été défini des stratégies de filtrage adaptées et consistantes. En particulier, cette thèse s'est focalisée sur les données qui peuvent être construites à partir du flux audio binaural, et sur la modélisation du lien existant entre ces observations audio (s'appuyant en pratique sur les indices interauraux traditionnels) et les variables de position et orientation entrant en jeux dans le problème. La stratégie de filtrage utilisée permet ainsi d'estimer, au fur et à mesure du mouvement, l'azimut et la distance entre la source et le robot, et de détecter conjointement si cette source est active ou non. En plus des développements théoriques, des expérimentations ont permis de démontrer l'applicabilité de l'approche dans un contexte réaliste, tel que celui représenté sur la figure 2.15a. Ces résultats ont été obtenu au cours d'une des visites d'A. PORTELLO à l'ISIR. Dans ces expérimentations, deux microphones sont placés sur un chariot mobile déplacé manuellement et au hasard. Un système de capture de mouvement fournit la vérité terrain ainsi que l'équivalent de la proprioception du robot. Sur cette base, on obtient la représentation de la figure 2.15b, où les positions absolues du récepteur binaural (bleu), du haut parleur fixe dans l'environnement (rouge) et l'estimation de position de l'algorithme de filtrage (gris) sont représentées.



(A) Détails du dispositif expérimental.



(B) Estimation de la position au fur et à mesure du mouvement.

FIGURE 2.15 – (A) Évaluation expérimentale de l'approche active. (Gauche) Aperçu du système, où deux microphones sont placés sur un chariot mobile, et un haut parleur (fixe) émet un son. Tous deux sont repérés dans l'espace via des LED infrarouges captées par un système de capture de mouvement. (Droite) Vue détaillée des microphones et du résultat de la capture 3D.
(B) Résultats expérimentaux en fonction du temps pour un capteur en champ libre (sans tête) et pour une source sonore émettant un signal vocal. Après un certain temps d'intégration du mouvement et des variation de sensations audio résultantes, l'approche permet de localiser très précisément la source dans son environnement. Figures tirées de (PORTELLO, 2013).

On peut y voir qu'au début de l'estimation réside une classique ambiguïté avant/arrière : la probabilité que la source sonore émette devant ou derrière le récepteur audio est la même et le mouvement effectué (ici une simple translation) ne permet pas de la lever. Ce n'est qu'une fois que le récepteur audio passe devant la source que cette ambiguïté n'existe plus, et la source sonore est alors très précisément localisée. Une des principales limites de ces travaux concerne le type de mouvement lui même : celui-ci est aléatoire et non contrôlé par le robot pour résoudre la tâche de localisation audio. La boucle perception/action traditionnelle n'est en fait pas fermée, et seule la voie montante, couplant les signaux audio extéroceptifs et la proprioception de la plateforme sont exploités pour inférer les positions relatives des sources sonores. Ces problématiques ont depuis été abordées par P. DANÈS dans (BUSTAMANTE, DANÉS et al., 2016).

Publication

Les travaux d'A. PORTELLO ont donné lieu aux articles (PORTELLO, DANÈS et ARGENTIERI, 2011 ; PORTELLO, DANÈS, ARGENTIERI et PLEDEL, 2013 ; MARKOVIC et al., 2013), tous trois publiés et présentés au sein des sessions spéciales "Robot Audition" de la conférence IEEE/RSJ IROS.

2.4 Vers des considérations attentionnelles audio-guidées

Les travaux présentés en 2.2, effectués durant la thèse de Karim YOUSSEF, étaient centrés sur l'estimation de la position d'une source sonore au sein d'un environnement acoustique réaliste. Nous avons pu montrer que si le robot était capable de s'y déplacer, alors il fallait tenir compte des changements acoustiques résultants pour espérer continuer à localiser avec une bonne précision les source sonores. En d'autres termes, l'action a été envisagée plutôt comme une contrainte venant s'appliquer à la tâche de localisation de source sonore. En parallèle, la thèse d'Alban PORTELLO brièvement évoquée en 2.3 a été l'occasion de traiter cette même problématique en exploitant cette fois les conséquences positives du changement de position, i.e. le changement de point de vue sur la ou les sources sonores au cours du mouvement. Par ailleurs, comme nous l'avons illustré à la figure 2.2, d'autres tâches audio sont susceptibles de bénéficier de la mobilité accordée à la plateforme mobile. Et si la localisation reste probablement la brique de base essentielle à tout système d'analyse d'une scène sonore en robotique, nous avons voulu également travailler à la place de l'action au sein d'une architecture complète impliquant l'ensemble des tâches de localisation, reconnaissance et navigation. Cette réflexion a été menée à l'occasion de la thèse de Benjamin COHEN-LHYVER, qui traite en particulier de la modulation du mouvement de tête d'un robot humanoïde pour l'élaboration de cartes multimodales permettant à l'agent de mieux comprendre son environnement. A la différence des travaux précédents, l'approche proposée par B. COHEN-LHYVER est intrinsèquement multimodale et active. Elle couple une approche montante traditionnelle d'analyse de signaux perceptifs (audio et vision) à une approche descendante permettant, via la génération d'ordres moteurs, de comprendre et interpréter l'environnement audiovisuel du robot.

Après une présentation du contexte et des objectifs de ces travaux dans une première sous-section, un système de modulation du mouvement de la tête (HTM pour Head Turning Modulation) est présenté. L'architecture de chacune des 2 parties le constituant (un module de pondération dynamique, jugeant de l'importante de la stimulation, couplé à un module

d'inférence et de fusion multimodale) est détaillée, et le système complet est enfin évalué dans une dernière sous-section. Une conclusion termine cette synthèse.

2.4.1 Positionnement et objectif

Imaginons un instant être étudiant, en cours. Alors que notre attention est (normalement) portée sur l'enseignant, un bruit de verre cassé au fond de la salle se fait entendre. Une des première réactions que nous aurons va être très certainement de tourner la tête en direction de l'événement audio, afin de comprendre ce qu'il s'est passé. Cette rotation de la tête permet de mobiliser la vision dans la direction estimée par la modalité auditive, et ainsi de porter notre attention sur cet événement inattendu. Ce mouvement reste néanmoins quasi-réflexe : l'analyse de l'événement perceptif n'a pas mobilisé de ressources cognitives avancées (aspect réflexe), mais reste sujet à une interprétation liée au contexte : nous avons reconnu un bruit de verre brisé, peu probable dans le contexte "salle de cours". Pour autant, si ce type de bruit se reproduit un certain nombre de fois, il est probable que nous ne tournerons plus la tête : l'événement à l'origine de la stimulation audio a été intégré à notre représentation de l'environnement. Et nous pouvons (normalement) porter à nouveau notre attention sur l'enseignant.

Doter un robot de ce type de capacité d'analyse de la scène (sonore, entre autre) reste aujourd'hui encore un défi. Dans l'exemple proposé plus haut, une solution pour décider du comportement actif du robot (via un déclenchement du mouvement de sa tête par exemple) pourrait consister à coder en dur ce comportement. Il s'agit alors d'envisager tous les cas possibles déclenchant une réaction du robot, et sur la base de règles codées par l'ingénieur, de décider du changement de son comportement. Évidemment, cela n'est guère envisageable dans des environnement dynamiques dotés d'objets audio et visuel changeant et inconnus a priori. D'une manière plus générale, chercher à comprendre les mécanismes à l'origine de la modulation attentionnelle et les reproduire dans un cadre robotique fait appel à de nombreux champs disciplinaires qu'il serait trop long de lister. Néanmoins, nous pouvons tenter de citer quelques inspirations à l'origine de ce travail, balayant des aspects biologiques, neuronaux, cognitifs et robotiques.

Dans notre contexte, nous restreindrons la définition de l'attention à la réquisition concomitante des organes sensibles (oreilles et yeux en particulier) vers une entité d'intérêt (spatialisée ou non). Très souvent, l'attention peut être déclenchée par les caractéristiques intrinsèques aux signaux captés : on parle alors d'un phénomène ascendant/montant, ou exogène (Le Meur et al., 2006 ; Driver et Spence, 1998), s'appuyant sur la *saillance* des caractéristiques bas-niveau des signaux (Treisman et Gelade, 1980). Il existe également des processus attentionnels descendant, ou endogène, motivant l'exploration spatiale ou sémantique dans un but déterminé et interne (Le Meur et al., 2006 ; Driver et Spence, 1998). La notion de saillance, bien qu'intensivement utilisée dans différentes communautés, reste assez mal définie, ou alors d'une manière spécifique à l'environnement et au contexte applicatif. Elle est néanmoins très souvent caractérisée comme une caractéristique *intrinsèque* à un stimuli pouvant être à l'origine d'une réaction attentionnelle, autant dans un cadre audio (Duangudom et Anderson, 2007) que visuel (Nothdurft, 2006). Cependant, dire qu'une telle caractéristique est susceptible de solliciter, a priori et systématiquement, l'attention du robot est réducteur. Comme l'a montré l'exemple illustratif précédent, le contexte, la répétition des phénomènes, sont susceptibles de moduler l'importance de cette saillance intrinsèque. Sur ce constat, nous avons proposé à l'occasion de la thèse de B. Cohen-Lhyver de définir une nouvelle forme de saillance, sémantique et contextualisée : *la congruence* (cf. §2.4.2.1). Ce type de considérations visant à inclure des informations contextuelles plus haut niveau dans l'établissement de cartes de saillance a déjà été proposé

dans la littérature (OLIVA et al., 2003), mais uniquement dans un cadre visuel et supervisé. Dans le domaine audio, nous pouvons citer (KAYSER et al., 2005 ; KALINLI et NARAYANAN, 2007). La prise en compte d'informations multimodales permet également l'émergence de comportements attentionnels basés sur la saillance d'une modalité en fonction d'une autre. Dans ce domaine, citons (RUESCH et al., 2008) qui propose un modèle de filtrage attentionnel basé sur la saillance de données multimodales via l'établissement d'une carte interne de l'environnement (PETERS II et al., 2001).

Il est intéressant de mettre en parallèle ces différents modèles de saillance avec les mécanismes neuronaux mis en place chez l'homme pour traiter de discontinuités sémantiques, i.e. d'événements imprédictibles. Cette détection est rendue possible par la capacité de nos aires sensorielles à détecter des stimuli inattendus. Une des manifestations neuronales connues de cette capacité est la MMN (Mismatch Negativity) (NÄÄTÄNEN, GAILLARD et MÄNTYSALO, 1978), qui se traduit par une augmentation de la réponse neuronale suite à la présentation d'un stimulus déviant (un son à 200Hz au milieu d'une séquences de sons à 1kHz, par exemple). Il est important de noter que la MMN se produira indépendamment de la signification propre du stimulus déviant ; en d'autres termes, la nature *incongrue* de la réaction dépend uniquement du contexte, et n'est pas issue d'une caractéristique intrinsèque au stimulus. Ce type de réaction a été identifié dans toutes les aires sensorielles du cerveau, mais est particulièrement présente dans le système auditif (MOLHOLM et al., 2005). Apparaissant entre 100ms et 200ms seulement après l'apparition du stimulus déviant, la MMN est également à l'origine de réactions motrices attentionnelles, illustrant ainsi l'intérêt de la notion de congruence dans la compréhension des entités perceptives peuplant notre environnement. Parmi les autres mécanismes d'intérêts, nous pouvons également citer la théorie de la hiérarchie inverse (RHT) (AHISSAR et HOCHSTEIN, 2004 ; NELKEN et AHISSAR, 2006 ; SHAMMA, 2008) qui stipule que le traitement multimodal d'informations sensorielles implique aussi bien une communication ascendante que descendante entre les capteurs et les aires computationnelles du cerveau. Cette théorie explique la propagation plus rapide des informations vers les aires de traitement haut-niveau lorsqu'il n'existe aucune ambiguïté entre les stimuli. Ainsi, l'analyse d'un son ou d'une image peut être accélérée par l'absence de temps passé à en analyser les composantes bas-niveau. A l'opposé, si cette même information est incongrue, il est alors nécessaire de mobiliser l'ensemble des aires de traitement, bas-niveau comme haut-niveau, pour en obtenir l'interprétation. Et la nature multimodale des stimuli renforce encore plus cette assertion, une modalité pouvant en compléter une autre.

Publication

Le positionnement et les considérations biologiques synthétisés dans cette sous section, tous deux peu connus et originaux au sein de la communauté Acoustique et Signal (binaural), ont donné lieu à la soumission par B. COHEN-LHYVER d'un chapitre du livre "The Technology of Binaural Understanding", édité par J. BLAUERT et J. BRAASCH. *Sa publication est prévue pour 2019.*

Doter un robot de la capacité de comprendre et structurer une représentation de son environnement, via la détection d'événements inattendus nécessitant son attention requiert ainsi (i) une définition précise de la notion d'incongruence (inspiré de la MMN), (ii) la nécessité de rationaliser les traitements en s'appuyant sur la nature multimodale des signaux à traiter (inspiré de la RHT), le tout (iii) en s'appuyant sur une architecture active capable d'exploiter les mouvements moteurs pour consolider cette représentation. Cette chaîne montante et descendante de traitement a justement été l'objet d'étude du projet Européen TWO!EARS, présenté dans la suite.

2.4.1.1 Le projet TWO!EARS

Le projet TWO!EARS a eu pour objectif de développer un modèle computationnel de la perception sonore (et de son expérience) intégrant des aspects multimodaux et actifs. Le projet a débouché en particulier sur la mise au point de 2 plateformes robotiques capables d'explorer de manière interactive leur environnement sur la bases de données audio (binaurales) et visuelles. Entre autres applications, le système logiciel et matériel TWO!EARS développé à cette occasion a pu servir de banc d'essai permettant l'évaluation de différents algorithmes mêlant une approche classique montante d'analyse des signaux audiovisuels à des considération descendantes inspirées de la cognition. Deux preuves de concept ont ainsi été proposées : l'analyse exploratoire d'une scène sonore –dans le contexte spécifique d'une tâche de recherche et sauvetage–, et la prédiction de la qualité de l'expérience basée sur l'exploration interactive de champs sonores pour l'évaluation de leur reproduction selon différentes méthodes. Le système repose sur une architecture modulaire ouverte [4] et largement documentée [5], ayant comme principale nouveauté la prise en compte explicite et imbriquée de processus d'analyse montant et/ou descendant. Enfin, une nouvelle approche formalisant la formation d'objets possiblement multimodaux est introduite, incluant l'affectation de leur sens, l'acquisition de connaissance, et la représentation (par apprentissage) de l'environnement. L'ensemble de ces points ont été traités par le consortium au sein d'une architecture particulière, présentée dans la suite.

2.4.1.2 Architecture du système TWO!EARS

La figure 2.16 représente schématiquement l'architecture complète du système TWO!EARS et met en évidence deux chemins de données :

— la voie montante : partant des signaux issus des capteurs (audio et visuels), une étape d'extraction de caractéristiques permet d'en extraire une représentation basse dimension. De nombreuses caractéristiques différentes ont été utilisées, allant des indices binauraux traditionnels présentés précédemment, aux traditionnels Mel Frequency Cepstral Coefficients (MFCCs) pour les méthodes de reconnaissance de sons. Pour les images, des gradients, surfaces, directions, sont extraites des images fournies par la paire stéréo équipant une des plateformes mobiles. Une originalité de l'architecture réside dans le fait que les paramètres d'extraction de ces caractéristiques peuvent pour la plupart être modifiés dynamiquement. Sur la base de ces caractéristiques, des systèmes experts proposent leur classification en terme de classes audio ou visuelle. Enfin, sur la base de ces classifications, une décision est prise quant à leur signification, ou sur la nécessité de modifier/moduler l'extraction de caractéristique pour affiner une conclusion. Cela nous mène au second chemin de données ;

— la voie descendante : les conclusions des experts (en terme de classification, de qualité de la localisation, etc.) peuvent modifier les paramètres des différentes couches traversées par les signaux. Il s'agit alors par exemple de modifier la répartition et la forme des filtres cochléaire (concentration sur une bande fréquentielle d'intérêt), de moduler le calcul de certaines caractéristiques (amplification/atténuation de certaines composantes du signal), d'amplifier ou amoindrir la reproduction artificielle de certains phénomènes psychoacoustiques (effet de précédence, pour le traitement des réverbérations par exemple). Cette voie descendante vient aussi possiblement moduler les commandes motrices du système, en particulier en provocant une modification du comportement exploratoire du robot.

4. L'intégralité des développements logiciels est disponible sur Github, à l'adresse https://github.com/TWOEARS/

5. http://docs.twoears.eu/en/1.4/

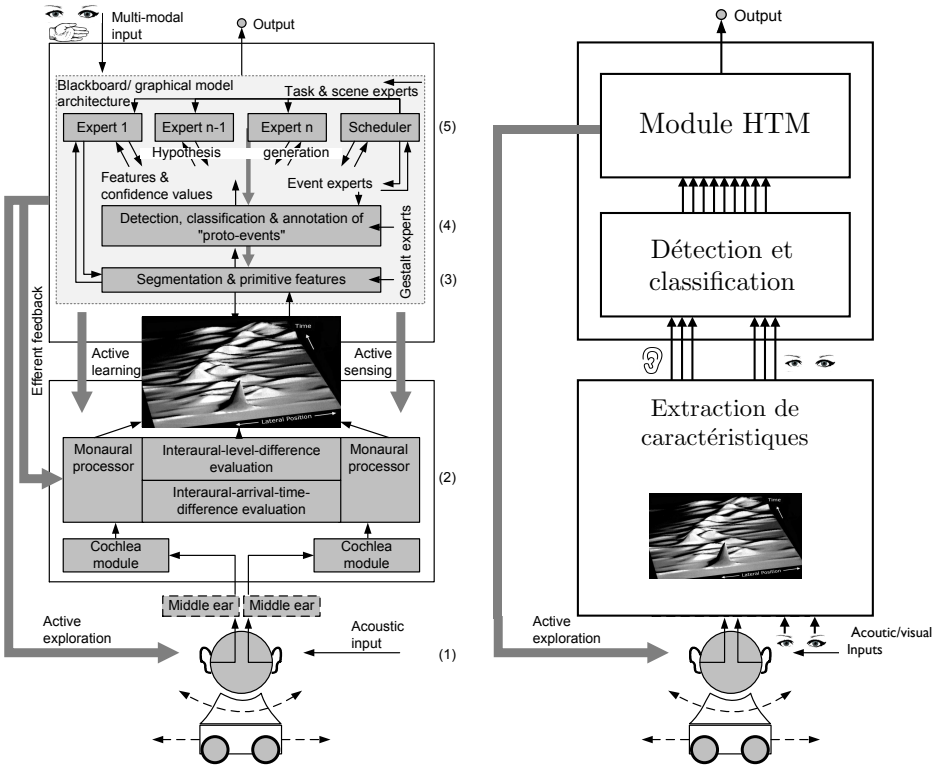


FIGURE 2.16 – (Gauche) Architecture complète du modèle TWO!EARS, telle que proposée dans le projet. (Droite) Version (très) simplifiée, précisant l'implémentation du module HTM prenant en charge la modulation du mouvement de la tête.

En terme d'implémentation, l'ensemble de ces fonctionnalités est structuré en trois principaux composants logiciels : un blackboard (structure de centralisation des données), des sources de connaissance (Knowledge Sources (KS), modules experts chargés de l'analyse précise et dédiée des données), et un scheduler (chef d'orchestre, responsable de l'exécution des différentes sources de connaissance).

Notre contribution au sein de cette architecture s'appuie principalement sur le retour moteur, permettant au système de se doter d'un comportement actif exploratoire, et en retour au robot de mieux comprendre et interpréter son environnement. Ainsi, l'architecture exploitée peut se "simplifier" en celle proposée à la figure 2.16 (droite), où le module "Head Turning Modulation" (HTM) de modulation du mouvement de la tête est exploité pour piloter le comportement moteur de la tête, tout en fournissant en sortie une interprétation de l'environnement du robot en terme d'objets multimodaux localisés. Le détail de cette approche est proposé dans la suite.

2.4.1.3 Notations

Comme indiqué au sein de la figure 2.16, le module HTM s'appuie sur la sortie des classifieurs audio et visuels. Un classifieur est une source de connaissance dédiée à la classification de trames audio ou visuelles en une seule et unique *classe* audio $\mathtt{c}_i^a$ ou visuelle $\mathtt{c}_k^v$ (par exemple, $\mathtt{c}_i^a \in \{\mathtt{voix}, \mathtt{aboiements}, \mathtt{cris}, \ldots\}$ pour les classes audio, ou $\mathtt{c}_k^v \in \{$PERSONNE, CHIENS, BÉBÉ, $\ldots\}$ pour les classes visuelles), de sorte que nous disposons d'autant de probabilité d'appartenance $p_i^a[t]$ et $p_k^v[t]$ d'une trame à une classe $\mathtt{c}_i^a$ et $\mathtt{c}_k^v$ qu'il y a de classifieurs audio ou visuel. Dans toute la suite, nous noterons respectivement $\mathbf{P}^a[t]$ et $\mathbf{P}^v[t]$ les vecteurs

regroupant les sorties des N_a et N_v classifieurs audio et visuels disponibles, de sorte que

$$\mathbf{P}^a[t] = (p^a_1[t], \ldots, p^a_{N_a}[t])^T \text{ et } \mathbf{P}^v[t] = (p^v_1[t], \ldots, p^v_{N_v}[t])^T, \tag{2.18}$$

où t représente l'indice temporel numérotant les trames catégorisées. De la même façon, nous disposons d'experts de localisation, fournissant pour chacun des angles potentiels d'arrivée (en azimut, le problème étant supposé plan) une probabilité de provenance $p^a_{\theta_u}[t]$ et $p^v_{\theta_u}[t]$. Par analogie, nous disposons respectivement donc de 2 vecteurs $\mathbf{\Theta}^a[t]$ et $\mathbf{\Theta}^v[t]$ de localisation audio et visuelle, définis par

$$\mathbf{\Theta}^a[t] = (p^a_{\theta_1}[t], \ldots, p^a_{\theta_{N_\theta}}[t])^T \text{ et } \mathbf{\Theta}^v[t] = (p^v_{\theta_1}[t], \ldots, p^v_{\theta_{N_\theta}}[t])^T, \tag{2.19}$$

avec $N_\theta = 72$ angles répartis entre 0° et 360° par pas de 5°. En pratique, l'ensemble de ces vecteurs sont regroupés en un seul et unique vecteur $\mathbf{V}[t]$, avec

$$\mathbf{V}[t] = (\mathbf{P}[t]^T, \mathbf{\Theta}[t]^T)^T, \text{ avec } \mathbf{P}[t] = (\mathbf{P}^a[t]^T, \mathbf{P}^v[t]^T)^T \text{ et } \mathbf{\Theta}[t] = (\mathbf{\Theta}^a[t]^T, \mathbf{\Theta}^v[t]^T)^T. \tag{2.20}$$

Ce vecteur $\mathbf{V}[t]$ constitue ainsi l'entrée du module HTM, dont l'objectif sera de transformer un des événements Ψ_j (apparition d'un son ou de données visuelles) présents de manière *objective* dans l'environnement $\epsilon^{(l)} = \{\Psi_1, \ldots, \Psi_{L_l}\}$, en un objet o_j perçu par le robot, i.e.

$$\Psi_j = \{\theta(\Psi_j), \mathsf{c}(\Psi_k)\} \longrightarrow o_j = \{\widehat{\theta}(o_j), \widehat{\mathsf{c}}(o_j)\}, \text{ avec } \widehat{\mathsf{c}}(o_j) = \{\widehat{\mathsf{c}}^a(o_j), \widehat{\mathsf{c}}^v(o_j)\}, \tag{2.21}$$

où un objet o_j (resp. un événement Ψ_j) est défini par sa localisation estimée $\widehat{\theta}(o_j)$ (resp. sa localisation $\theta(\Psi_j)$), mais également par sa *classe audiovisuelle estimée* $\widehat{\mathsf{c}}(o_j)$ (resp. sa classe audio visuelle $\mathsf{c}(\Psi_k)$). A noter que l'estimation de la position angulaire d'un objet s'effectue sur la base de l'information visuelle tant que l'objet est dans le champ visuel du robot ; dans le cas contraire, c'est la localisation audio qui est exploitée. Cette définition générique permet alors d'envisager des cas courants où l'objet perçu o_j ne coïncide pas avec l'événement Ψ_j objectivement présent dans l'environnement : des erreurs de localisation ou de reconnaissance peuvent en effet dégrader la représentation interne $e^{(l)} = \{o_1, \ldots, o_{N_l}\}$ du $l^{\text{ième}}$ environnement analysé par le robot, définie par la collection des N_l objets la constituant. Enfin, nous pouvons définir les catégories audiovisuelles $\mathcal{C}^{(l)}(\mathsf{c}^a_i, \mathsf{c}^v_k)$ présentes dans cette représentation par

$$\mathcal{C}^{(l)}(\mathsf{c}^a_i, \mathsf{c}^v_k) = \left\{ o_j \in e^{(l)}, \widehat{\mathsf{c}}^a(o_j) = \mathsf{c}^a_i \text{ et } \widehat{\mathsf{c}}^v(o_j) = \mathsf{c}^v_k \right\}. \tag{2.22}$$

Sur la base de ces définitions, comment obtenir et consolider au fur et à mesure de l'exploration une représentation interne stable de l'environnement de manière active ? C'est à cette question que tente de répondre l'architecture de modulation du mouvement de la tête présentée dans la suite.

2.4.2 Le module HTM

Une des première étapes du travail de thèse de B. COHEN-LYVER a consisté à travailler sur la notion d'importance. Cette première réflexion a donné lieu à un module, au sein de l'architecture du projet TWO!EARS, dit de "pondération dynamique" et visant à déclencher des mouvements de la tête en direction des objets détectés comme étant importants pour la compréhension de l'environnement. Cette importance est formalisée via la notion de congruence explicitée précédemment. Puis, très vite s'est posé la question de la multimodalité et de l'accès aux données audio et visuelle dans un contexte où le robot pouvait être amené à focaliser ses capteurs en direction d'une source d'intérêt. Alors un second module, dit "d'inférence et de fusion multimodale" et visant à compléter une information éventuellement manquante sur un objet audiovisuel sur la seule base des données perçues, a été proposé par B. COHEN-LYVER. Il s'agit alors d'être capable d'associer, par apprentissage actif, les labels audio et visuel des événements présents autour du robot. Sur cette base, une représentation interne de l'environnement peut enfin être construite : le robot dispose alors de toute l'information nécessaire pour moduler son comportement. Ces deux modules sont rapidement présentés et évalués dans la suite.

2.4.2.1 Module de pondération dynamique

Le module de pondération dynamique, appelé DWmod dans la suite, vise à implémenter la notion de congruence de façon à décider de l'importance d'un objet dans la représentation interne de l'environnement d'un robot. A la différence des approches mentionnées en introduction, ce qui doit requérir l'attention du robot ne doit pas être le signal dont les caractéristiques bas-niveau forment un marqueur de différence temporel (via la notion de saillance), mais plutôt celui dont le sens est marqueur de différence par rapport à son environnement sémantique. De fait, l'approche proposée diffère des travaux déjà existant selon :

— le contenu audio et visuel bas niveau, qui n'est donc pas pris en compte. Seule son interprétation haut niveau, i.e. sa catégorie audiovisuelle, est pertinente ;

— le rôle du comportement actif, qui va d'une part permettre au système d'apprendre son environnement, mais qui va également devoir être modulé selon son degré de compréhension.

Formalisation : Dans toute la suite, un objet o_j de classe audio c_i^a et de classe visuelle c_k^v sera supposé *incongru* si d'autres objets appartenant à la même catégorie audiovisuelle $\mathcal{C}^{(l)}(\mathsf{c}_i^a, \mathsf{c}_k^v)$ n'ont jamais été détectés précédemment par le robot. Sur la base de cette définition, et en omettant temporairement la dépendance temporelle des notations, nous pouvons définir la probabilité $p(\mathcal{C}^{(l)}(\mathsf{c}_i^a, \mathsf{c}_k^v))$ qu'un objet o_j appartienne à la catégorie audiovisuelle $\mathcal{C}^{(l)}(\mathsf{c}_i^a, \mathsf{c}_k^v)$ à un instant t par

$$p(\mathcal{C}^{(l)}(\mathsf{c}_i^a, \mathsf{c}_k^v)) = \frac{|\mathcal{C}^{(l)}(\mathsf{c}_i^a, \mathsf{c}_k^v)|}{N_l}, \tag{2.23}$$

où la notation $|.|$ désigne le cardinal des ensembles. Sur la base de cette probabilité, nous pouvons alors définir une pondération $w(o_j)$ de l'objet o_j selon

$$w(o_j) = \begin{cases} \overline{f}[n] \text{ si } p(\mathcal{C}^{(l)}(\mathsf{c}_i^a, \mathsf{c}_k^v)) \leq K_l, \\ \underline{f}[n] \text{ sinon,} \end{cases} \tag{2.24}$$

où $K_l = 1/|\mathcal{C}^{(l)}|$ représente l'équiprobabilité des catégories audiovisuelles, et n le nombre de trames durant lequel sa probabilité d'appartenance à la catégorie $\mathcal{C}^{(l)}(\mathsf{c}_i^a, \mathsf{c}_k^v)$ a été inférieur ou supérieur à K_l [6]. Cette valeur de seuil a été choisie afin de n'imposer aucune donnée a priori, évitant ainsi l'apparition de biais dans les fréquences attendues d'apparition des catégories audiovisuelles. Les deux fonctions de pondération $\overline{f}$ et $\underline{f}$ sont ici deux sigmoïdes au comportement symétrique, $\overline{f}$ partant de 0 pour tendre vers 1 en $n = 5$ pas de temps (et symétriquement pour $\underline{f}$). Ainsi, un objet o_j sera considéré comme incongru (et donc il sera pertinent d'y porter son attention) si son poids est égal à 1, i.e. si sa probabilité d'appartenir à la catégorie audiovisuelle $\mathcal{C}^{(l)}(\mathsf{c}_i^a, \mathsf{c}_k^v)$ est inférieure ou égale à l'équiprobabilité pendant au moins $n = 5$ pas de temps.

Sur la base de cette décision sur l'incongruence éventuelle d'un objet, il faut alors décider du déclenchement du mouvement de la tête en sa direction. Or plusieurs événements peuvent être considérés comme incongrus au même moment. Il est donc nécessaire de sélectionner l'objet nécessitant l'attention du robot, c'est à dire celui qui provoquera une rotation de la tête en sa direction. Si on note $\Delta t(o_j)$ la durée durant laquelle un objet o_j est présent dans la représentation de l'environnement, on peut alors définir le vecteur d'activité τ_{DW} des N_l objets présents dans cette même représentation par

$$\tau_{\text{DW}} = \{\tau_{\text{DW}}(o_1), \ldots, \tau_{\text{DW}}(o_{N_l})\}, \text{ avec } \tau_{\text{DW}}(o_j) = \Delta t(o_j) \times \frac{p(\mathcal{C}^{(l)}(\mathsf{c}_i^a, \mathsf{c}_k^v))}{K_l}. \tag{2.25}$$

Ainsi, disposant de chacune des localisations estimées $\widehat{\theta}^{a/v}(o_j)$ des N_l objets, on décide de générer la commande motrice θ_{DW} selon

$$\theta_{\text{DW}} = \widehat{\theta}(o_j) \text{ avec } j = \arg\min_l \tau_{\text{DW}}(o_l). \tag{2.26}$$

Ainsi, l'objet o_j disposant de l'activité $\tau_{\text{DW}}(o_j)$ la plus faible est choisi comme étant celui vers lequel le robot devra tourner la tête. Il est intéressant de noter que l'équation (2.25) privilégie les objets venant d'apparaître dans la représentation, associés à une valeur $\Delta t(o_j)$ faible. Cette modulation par la durée d'apparition permet ainsi d'inclure une forme de motivation par la nouveauté, et donc une certaine forme de curiosité, dans le comportement du robot. Au final, dans le cas extrême où chaque nouvel objet apparaissant dans la représentation appartient à une nouvelle catégorie audiovisuelle, le seuil K_l (fonction de l'équiprobabilité d'apparition de ces catégories) tend à diminuer de plus en plus, provoquant ainsi un mouvement de tête systématique dans leur direction. A l'opposé, si la quasi-totalité des objets appartiennent à la même catégorie, K_l augmentera de sorte que la plupart des mouvements de tête seront inhibés : l'apparition de sources identiques ne nécessite pas l'attention du robot, libérant ainsi ses mouvements pour d'autres tâches.

Pour terminer il est clair que pour une représentation $e(l)$ d'un environnement donné (définie pour rappel par la collection des objets o_j la constituant) les poids associés à chacun de ses objets constituent une distribution de leurs catégories. En ce sens, le vecteur $\mathbf{W}^{(l)}$ défini par

$$\mathbf{W}^{(l)} = \{w(o_1), \ldots, w(o_J)\} \tag{2.27}$$

représente les *règles de congruence* du $l^{\text{ième}}$ environnement représenté. Ainsi, à chaque nouvel environnement exploré peut correspondre des règles de congruence différentes. Néanmoins, garder en mémoire ces règles de congruence permet éventuellement d'en hériter lors de l'exploration d'un environnement inconnu. Il s'agit alors de *transférer* les connaissances

6. En pratique, n est remis immédiatement à 0 en cas de changement de décision sur la congruence ou incongruence de l'objet.

acquises précédemment vers la nouvelle représentation : un objet déjà congru dans un environnement précédent peut ainsi directement être considéré comme tel dans un nouveau. Nous pouvons donc appliquer les règles de congruence $\mathbf{W}^{(n)}$ de la représentation $e^{(n)}$ à la représentation $e^{(m)}$ en cours d'exploration si et seulement si

$$\{\mathcal{C}^{(m)}(\mathsf{c}_i^a, \mathsf{c}_k^v)\} \subseteq \{\mathcal{C}^{(n)}(\mathsf{c}_i^a, \mathsf{c}_k^v)\}, \tag{2.28}$$

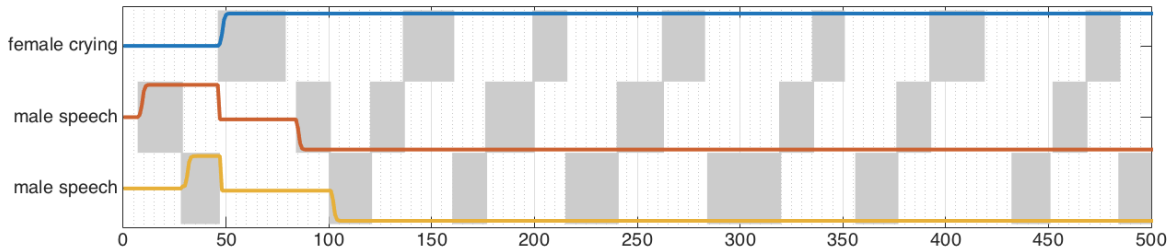
et nous aurons alors dans ce cas $\mathbf{W}^{(m)} = \mathbf{W}^{(n)}$: les règles de congruence déjà découvertes ont été transférées à un nouvel environnement.

Évaluation : Nous pouvons d'ores et déjà illustrer le fonctionnement du DWmod dans le cadre d'un scénario simulé simple, impliquant uniquement des objets audiovisuels émettant chacun leur tour dans le temps. Des évaluations en (bien) plus grand nombre sont proposées dans (COHEN-LHYVER, 2017), notamment dans un cas multisources simultanées. La figure 2.17a représente l'évolution dans le temps des poids $w(o_j)$ pour une scène constituée de 3 objets : 2 de catégorie identique (MALE `speech`), et 1 de catégorie différente (FEMALE `crying`). A l'apparition du premier objet, celui-ci se voit immédiatement attribué un poids de 1 (il est donc incongru). L'apparition d'un second objet d'une catégorie identique ne change pas la classification des objets en présence (les premières sources apparaissant dans l'environnement sont sujettes à un cas limite imposé par l'égalité non-stricte dans (2.24)). L'apparition d'un nouvel objet d'une nouvelle catégorie à $t \approx 50$ conduit à une réévaluation de leurs poids, qui seront à nouveau réévalués à l'occasion de leur réapparition en $t = 80$ et $t = 100$: les deux sources deviennent alors congrues. Jusqu'à la fin de la simulation, aucun nouvel objet n'apparaissant dans l'environnement, c'est donc l'objet FEMALE `crying` qui reste incongru. On peut noter que même si d'autres objets MALE `speech` étaient apparus, ils auraient été immédiatement classés comme congrus : dans cet environnement, cet objet n'est pas porteur d'information, et il n'est vraisemblablement pas utile que le robot tourne sa tête en leur direction (sauf pour une tâche autre que celle envisagée ici).

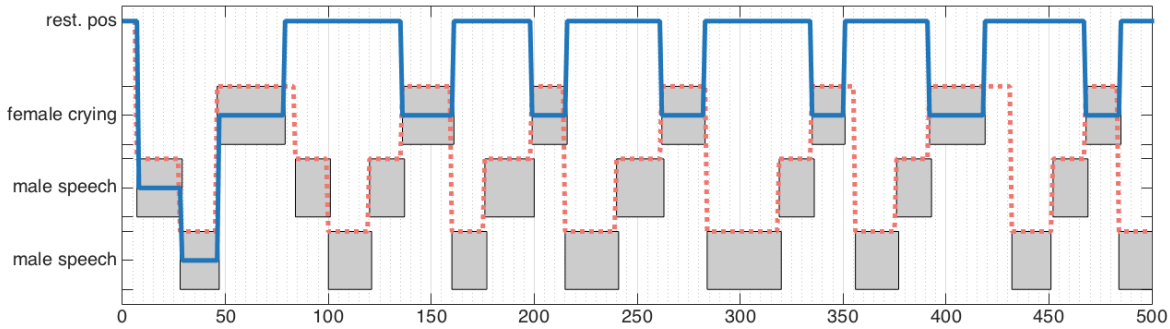
Les conséquences sur les mouvements de tête, et donc le comportement moteur exploratoire du robot, sont évidentes. La figure 2.17b indique qu'après s'être focalisé sur les deux premiers objets MALE `speech` présents dans l'environnement, le robot se focalise systématiquement sur l'objet incongru FEMALE `crying` (cf. tracé en trait plein). En comparaison, un robot naïf qui tournerait la tête en réponse à toutes les sollicitations de l'environnement serait nettement plus sollicité (cf. tracé en pointillés). Le DWmod a donc modulé le réflexe de rotation de la tête en ayant intégré à sa représentation les fréquences de présentation des objets la constituant. Nous voyons néanmoins apparaître une des limites de la formalisation simple employée ici : aucune habituation temporelle autre que celle basée sur les fréquences d'apparitions des objets n'est exploitée, de sorte que le robot reste toujours focalisé sur l'objet incongru. Il serait néanmoins très simple d'ajouter de telles considérations, mais qui peuvent alors êtres envisagées à l'occasion d'une tâche de plus haut-niveau (décisionnelle en particulier), non traitée ici.

Publication

La contribution présentée dans cette sous-section, détaillant la formalisation du module de pondération dynamique, a donné lieu à l'article (COHEN-LHYVER, ARGENTIERI et GAS, 2015) publié et présenté au sein de la conférence IEEE ROBIO.



(A) Évolution des poids associés à chacun des objets. A un poids négatif correspond un objet congru, à un poids positif correspond un objet incongru.



(B) Objets focalisés par le robot. Traits bleus : trajectoire motrice générée par le DWmod (un objet "focalisé" est traversé par cette ligne) ; traits rouges : trajectoire motrice obtenue pour un robot naïf.

FIGURE 2.17 – Représentation du fonctionnement du DWmod pour différents objets qui "émettent" leur information audiovisuelle aux instants repérés en gris. Figure tirée de (COHEN-LHYVER, 2017).

2.4.2.2 Module d'inférence et de fusion multimodale

Le module DW avait pour objectif de déterminer quel(s) événement(s) audiovisuel(s) pourrai(en)t nécessiter l'attention du robot dans un contexte d'exploration d'environnements inconnus. S'appuyant sur la notion de congruence, vue comme une mesure de l'importance sémantique relative d'un événement audiovisuel, le système proposé dans la première partie de la thèse de B. COHEN-LHYVER permet de doter le robot d'une représentation de son environnement. En ce sens, il est à son tour un "système expert" (au même titre que les classifieurs audio et visuel), partageant cette représentation du monde au sein du blackboard de l'architecture TWO!EARS. Cependant, on remarque assez vite un soucis majeur lié à la façon dont le module a accès aux données (ici les résultats de classification). Nous avons en effet fait l'hypothèse que les classes audiovisuelles étaient systématiquement accessibles au robot, alors qu'il est évident que ce ne sera pas le cas : si une source est placée en dehors du champs de vision des caméras, le label visuel sera inaccessible. Dès lors, comment décider de la congruence d'un événement ? Les résultats précédents ont été obtenus en faisant l'hypothèse que toutes les données étaient accessibles à tout instant (via une vision omnidirectionnelle par exemple). En pratique, un module supplémentaire d'inférence et de fusion multimodale doit être mis en œuvre afin de compléter les éventuelles données manquantes (typiquement, une source non visible). Et nous avons proposé à l'occasion de la thèse de B. COHEN-LYVER que ce module supplémentaire exploite à son tour les capacités d'action du robot pour sa tâche. C'est ce module qui est présenté dans la suite .

Le module d'inférence et de fusion multimodale (module MFI) est placé en amont du module DW. Il a pour objectif de fournir les classes audiovisuelles estimées $\widehat{c}(o_j)$ des objets présents dans l'environnement du robot, et ce :

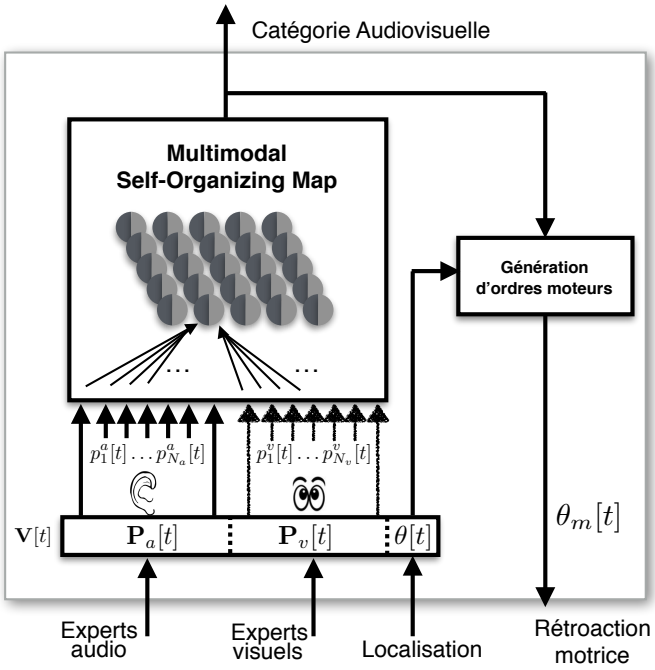


FIGURE 2.18 – Architecture globale du module MFI. Les sorties des classifieurs audio et visuel sont utilisées en entrée d'une carte auto-organisatrice fusionnant les deux modalités. Son apprentissage nécessite de tourner la tête en direction des événements perçus : l'approche est donc active, et le mouvement est ici utilisé pour compléter les données manquantes, ou les confirmer et cas de prédiction non suffisamment certaine.

— même si la donnée audio ou visuelle est manquante (événement non vu par les caméras, ou non entendu par les microphones),

— malgré la présence inévitable d'erreurs de classification des classifieurs audio et visuel.

Pour ce faire, le module MFI devra apprendre le lien entre les classes audio et visuelle (c'est ainsi que par exemple, reconnaître une vache permet de prédire le son qu'elle produit, et vice-versa), mais également corriger les sorties des classifieurs en cas de mauvaise classification de la trame audiovisuelle. L'architecture du module est représentée à la figure 2.18. Elle met en évidence la particularité du système proposé : alors que le lien audiovisuel peut être appris sur la base d'exemples de paires de classes audiovisuelles fournies en amont, nous proposons ici à nouveau une approche ne requérant aucun a priori sur la scène à analyser, et exploitant l'action (le mouvement de la tête) pour découvrir les liens existant entre les modalités. La figure 2.18 montre ainsi une voix montante d'analyse, s'appuyant sur un réseau de neurones de type "carte auto-organisatrice" (SOM) fusionnant les modalités audio et visuelle, et une voix descendante générant une réaction motrice en direction des événements audiovisuels à apprendre. Ce comportement moteur est une des contributions fortes de l'architecture, peu d'approche envisageant d'aller chercher, par l'action, les données manquantes nécessaires à l'apprentissage. Ces deux voies de traitement sont formalisées dans la suite.

Formalisation : Nous proposons d'utiliser ici un SOM pour apprendre d'une manière non supervisée le lien entre les catégories audio et visuelle définissant un objet dans la représentation de l'environnement. Ce SOM sera utilisé en ligne, c'est à dire qu'il recevra, à chaque pas de temps t, une trame audiovisuelle $\mathbf{V}[t]$ définie selon (2.20). Il est clair qu'une structure de SOM classique n'est pas adaptée à la résolution de notre problème : si une modalité est manquante, il n'est pas possible d'ignorer les sorties des classifieurs concernés (le vecteur de données serait alors de taille variable), ni de mettre arbitrairement à 0 leurs sorties

(ces 0 arbitraires ayant une signification dans l'espace des données). En conséquence, nous avons introduit une nouvelle structure de carte auto-organisatrice appelée M-SOM (Multimodal Self- Organizing Map), composée de deux cartes audio et visuelle à deux dimensions, toutes deux composées de $I \times J$ nœuds, notés $r_{i,j}$ avec i, j la position (identique) du nœud dans les deux cartes. A noter qu'un nœud possède une certaine connectivité avec ses voisins (le cas des nœuds au bord de la carte étant particulier). A chaque nœud est associé deux vecteurs de poids $\mathbf{w}_{i,j}^{a/v} = \left(w_{i,j}^{a/v}(1), \ldots, w_{i,j}^{a/v}(N_a + N_v)\right)^T$ de même dimension que les données d'entrée (la notation a/v représente un raccourci pour représenter la grandeur audio ou visuelle de manière compacte). Ces données sont regroupées au sein d'une matrice $\mathbf{P}$ de taille $M \times N$, résultat de la concaténation des N vecteurs $\mathbf{V}[t]$ de taille M à l'instant t. Après initialisation des poids, les différentes étapes de l'approche sont les suivantes.

Si toutes les modalités sont disponibles : dans ce cas, les classifieurs audio et visuel fournissent des résultats de classification, i.e. des probabilités d'appartenance aux classes qu'ils représentent regroupées au sein des vecteurs $\mathbf{P}^a$ et $\mathbf{P}^v$ définis par (2.18). Le nœud r_{BMU}^{av} représentant le mieux ces données audiovisuelles est alors déterminé selon

$$r_{\text{BMU}}^{av} = r_{I,J}, \text{ avec } I, J = \arg\min_{i,j} \left(\|\mathbf{P}^a - \mathbf{w}_{i,j}^a\| \times \|\mathbf{P}^v - \mathbf{w}_{i,j}^v\| \right). \tag{2.29}$$

Ainsi, r_{BMU}^{av} est le nœud dont le vecteur de poids $\mathbf{w}_{\text{BMU}}^{av} = (\mathbf{w}_{\text{BMU}}^a, \mathbf{w}_{\text{BMU}}^v)$ est composé des vecteurs de poids audio *et* visuel les plus similaires, au sens de la distance euclidienne, aux vecteurs d'entrée $\mathbf{P}^a$ et $\mathbf{P}^v$. Une fois ce nœud vainqueur déterminé, les vecteurs de poids des deux cartes audio et visuelle sont mis à jour à l'aide d'une fonction de voisinage $h_{i,j}$ de façon à propager l'apprentissage aux voisins, selon la récurrence

$$\begin{aligned} &\mathbf{w}_{i,j}^{a/v}[n+1] = \mathbf{w}_{i,j}^{a/v}[n] + \alpha[n] h_{i,j}[n] \|\mathbf{P}^{a/v} - \mathbf{w}_{i,j}[n]\|, \\ &\text{avec } h_{i,j}[n] = \exp\left(\frac{\|r_{\text{BMU}}^{av}[n] - r_{i,j}[n]\|^2}{2\sigma[n]} \right), \end{aligned} \tag{2.30}$$

où n représente l'indice d'itération de la phase d'apprentissage, et σ la variance de la fonction Gaussienne de voisinage dont la valeur précise l'amplitude et la taille dans la carte de la propagation de l'apprentissage aux nœuds voisins. Une fois l'étape d'apprentissage terminée, la détermination du nœud vainqueur r_{BMU}^{av} permet d'estimer la classe audiovisuelle de la trame en cours d'analyse, et donc de l'objet y correspondant, par

$$\widehat{\mathtt{c}}^{\text{all}}(o_j) = \{\widehat{\mathtt{c}}^a(o_j), \widehat{\mathtt{c}}^v(o_j)\} = \{\mathtt{c}_i^a, \mathtt{c}_k^v\}, \text{ avec } i = \arg\max_l w_{\text{BMU}}^a(l) \text{ et } k = \arg\max_m w_{\text{BMU}}^v(m), \tag{2.31}$$

où l'exposant $^{\text{all}}$ précise que les deux modalités étaient disponibles pour décider de la classe audiovisuelle. On constate donc que cette décision s'opère sur un vecteur de poids du M-SOM, et non plus directement en sortie des classifieurs. Le M-SOM agit donc comme un filtre, qui grâce à la phase d'apprentissage mentionnée précédemment, est susceptible de *corriger* les erreurs de classification initiales. Ce point sera spécifiquement évalué plus loin dans le document.

Si une des modalités est manquante : supposons maintenant qu'une des modalités audio ou visuelle est manquante (l'objet n'émet pas de son, il est occulté, etc.). Dans une telle situation, l'apprentissage mentionné précédemment ne peut avoir lieu. Par contre, les étapes d'apprentissage précédentes peuvent être exploitées pour inférer la classe audio ou visuelle manquante. Prenons, sans perte de généralité, le cas où une donnée visuelle est manquante (cas classique de l'événement en dehors du champ visuel du robot). Alors,

1. l'audio seul est utilisé pour déterminer le nœud vainqueur r^a_{BMU} dans la carte audio, dont le vecteur de poids associé $\mathbf{w}^a_{\text{BMU}}$ permet de déterminer la classe audio $\widehat{\mathtt{c}}^a(o_j) = \mathtt{c}^a_i$, avec $i = \arg\max_l w^a_{\text{BMU}}(l)$;
2. le nœud visuel gagnant est directement hérité du nœud audio selon $r^v_{\text{BMU}} = r^a_{\text{BMU}}$: c'est précisément à cette étape que l'apprentissage du lien entre les classes audio est visuelle est exploité ;
3. au nœud visuel gagnant est associé le vecteur de poids $\mathbf{w}^v_{\text{BMU}}$, qui permet à son tour de déterminer la classe visuelle $\widehat{\mathtt{c}}^v(o_j) = \mathtt{c}^v_k$, avec $k = \arg\max_m w^v_{\text{BMU}}(m)$;
4. ainsi, la classe audiovisuelle estimée $\widehat{\mathtt{c}}^{\text{miss}}(o_j)$ de l'objet analysé à l'instant t est alors donnée par $\widehat{\mathtt{c}}^{\text{miss}}(o_j) = \{\widehat{\mathtt{c}}^a(o_j), \widehat{\mathtt{c}}^v(o_j)\}$, où l'exposant $^{\text{miss}}$ souligne le fait que la catégorie estimée a été obtenue sur la base de données manquantes.

Bien entendu, le raisonnement est identique dans le cas où la modalité audio seule est manquante : le résultat de classification visuel est utilisé pour estimer la classe audio correspondante, et donc la classe audiovisuelle complète.

Il est clair qu'en début d'apprentissage, l'inférence de la classe manquante a de forte chance d'être erronée. Le lien entre les classes audio et visuelle n'est en effet pas encore bien établi, dans la mesure où il ne peut s'opérer que lorsque les deux modalités sont présentes. Pourtant, les données complètes sont très souvent disponibles dans l'environnement : par exemple, en cas de donnée visuelle manquante, il suffit très souvent que le robot tourne la tête en direction de la stimulation audio pour retrouver les données complètes, et effectuer une nouvelle passe d'apprentissage. C'est précisément ce comportement que nous avons implémenté, via la définition d'un *ratio d'inférence* q indiquant si les classes audiovisuelles ont été correctement apprises (i.e. nous pouvons faire confiance en l'inférence) ou non (auquel cas il faudra confirmer l'inférence par une recherche motrice de la donnée manquante). Ce ratio d'inférence est défini par

$$q\left(\mathcal{C}^{(l)}(\mathtt{c}^a_i, \mathtt{c}^v_k)\right)[t] = \frac{\sum_1^t \delta^{\text{miss}}_{ik}[t-1]\delta^{\text{all}}_{ik}[t]}{\sum_1^t \delta^{\text{miss}}_{ik}[t]}, \text{ avec } \delta^{\text{all/miss}}_{ik}[t] = \begin{cases} 1 \text{ si } \widehat{\mathtt{c}}^{\text{all/miss}}(o_j) = \{\mathtt{c}^a_i, \mathtt{c}^v_k\}, \\ 0 \text{ sinon.} \end{cases} \tag{2.32}$$

Ainsi, le ratio d'inférence q pour la classe audiovisuelle $\{\mathtt{c}^a_i, \mathtt{c}^v_k\}$ est le rapport entre le nombre de fois que cette classe a été estimée sur la base de données manquantes en $t-1$ et confirmée en t, sur le nombre total d'inférences. La confirmation d'une inférence effectuée en $t-1$ dans le cas d'une des modalités manquantes est effectuée via un mouvement de la tête en direction de la source d'intérêt. Il est attendu que plus l'apprentissage du lien audiovisuel sera avancé, plus l'inférence sera confirmée par la recherche de la donnée manquante, de sorte que q tende vers 1. Ainsi, la comparaison du ratio q avec un seuil K_q permet de décider à partir de quand nous faisons confiance à l'inférence : il ne sera dès lors plus nécessaire d'effectuer un mouvement de tête, le robot ayant confiance en son inférence. Ce seuil K_q permet ainsi de moduler le comportement moteur du robot. De manière analogue à (2.25), nous pouvons définir un vecteur d'activité τ_{MFI} des N_l objets présents par

$$\tau_{\text{MFI}} = \{\tau_{\text{MFI}}(o_1), \ldots, \tau_{\text{MFI}}(o_{N_l})\}, \text{ avec } \tau_{\text{MFI}}(o_j) = \delta[n] \times \frac{q\left(\mathcal{C}^{(l)}(\mathtt{c}^a_i, \mathtt{c}^v_k)\right)}{K_q}, \tag{2.33}$$

où $\delta[n]$ représente une fonction de pondération visant à introduire une forme de persistance dans le comportement moteur de sorte à éviter la génération de mouvements contradictoires à chaque pas de temps. Ainsi, disposant de chacune des localisations audio et/ou visuelle $\theta^{a/v}_j$ des N_l objets, on décide de générer la commande motrice θ_{MFI} selon

$$\theta_{\text{MFI}} = \widehat{\theta}(o_j) \text{ avec } j = \arg\min_l \tau_{\text{MFI}}(o_l). \tag{2.34}$$

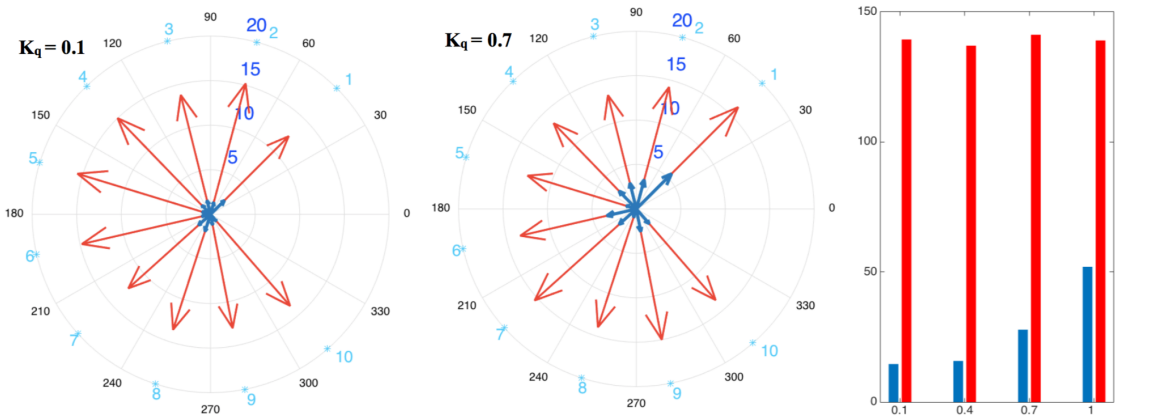
ϵ	N_l	MFI	Robot omniscient	Robot naïf	Ratio
0.3	3	0.992 (0.020)	0.703 (0.042)	0.414 (0.055)	1.41
	5	0.987 (0.022)	0.692 (0.017)	0.265 (0.014)	1.43
	7	0.942 (0.028)	0.691 (0.014)	0.198 (0.017)	1.36
	10	0.883 (0.041)	0.689 (0.011)	0.0145 (0.014)	1.28
	moy	**0.951**	**0.693**	**0.255**	**1.37**
0.7	3	0.774 (0.087)	0.282 (0.030)	0.165 (0.028)	2.74
	5	0.737 (0.105)	0.294 (0.014)	0.120 (0.023)	2.5
	7	0.683 (0.133)	0.296 (0.016)	0.081 (0.012)	2.31
	10	0.550 (0.117)	0.293 (0.016)	0.064 (0.011)	1.88
	moy	**0.686**	**0.291**	**0.107**	**2.36**

TABLE 2.1 – Taux de bonne estimation de la catégorie audiovisuelle (variance entre parenthèses). ϵ représente le taux d'erreur par trame des classifieurs, N_l le nombre de sources présentes dans l'environnement (et émettant possiblement en même temps).

Ainsi, l'objet o_j disposant de l'activité $\tau_{\text{MFI}}(o_j)$ la plus faible est choisi comme étant celui vers lequel le robot devra tourner la tête pour renforcer l'apprentissage de son lien audiovisuel.

Évaluation : Nous pouvons illustrer le bon fonctionnement du module de fusion et inférence à l'occasion de simulations permettant dans un premier temps d'évaluer sa capacité à estimer les bonnes classes audiovisuelles des objets présents dans la scène. A nouveau, cette évaluation s'effectue dans un cadre artificiel, où les sorties des classifieurs (pour rappel, fournissant une probabilité d'appartenance à la classe audio ou visuelle qu'ils représentent individuellement) sont simulées. Ces classifieurs sont supposés non idéaux, et leurs sorties sont soumises à un taux d'erreur ϵ pour chaque trame (supposée indépendante). Les performances du système sont comparées à un système de fusion naïf, disposant de 2 stratégies d'estimation des classes audiovisuelles : soit le robot a accès en permanence aux données audio et visuelle (le robot est omniscient), soit le système ne dispose que des données présentes en face de lui, et réagit en tournant la tête à l'apparition de chacune des sources de l'environnement (le robot est naïf). Dans les 2 cas, la catégorie audiovisuelle de la trame est estimée en sélectionnant les classes audio et visuelles possédant la probabilité la plus forte en sortie des classifieurs : aucune fusion n'est alors effectuée. 5 simulations des mêmes environnements (définis par la collection des N_l objets les constituant) est effectuée de façon à obtenir le taux moyen de bonne estimation de la catégorie audiovisuelle, reporté au sein du tableau 2.1. Les résultats montrent que même dans le cas de classifieurs soumis à des taux d'erreurs très importants (jusqu'à $\epsilon = 70\%$) le module MFI propose une estimation des classes audiovisuelles plus de deux fois meilleurs que dans le cas d'une fusion naïve effectuée en présence de la totalité des données (ce qui en pratique n'arrive jamais). En comparaison avec un robot totalement naïf récupérant les données uniquement à l'occasion d'un mouvement de tête en direction de la dernière source active, le module MFI propose des estimations prêt de 9 fois meilleures. La stratégie active du MFI, allant chercher par le mouvement les données manquantes pour apprendre le lien entre les classes audio et visuelles des objets, démontre ainsi sa grande efficacité.

Une autre illustration de son fonctionnement consiste à examiner le nombre de mouvements de tête générés pour l'apprentissage du lien audiovisuel en comparaison avec le robot naïf mentionné précédemment. Une telle étude est proposée à la figure 2.19 où sont représentés deux comportements moteurs, obtenus pour deux valeurs différentes du seuil $K_q < 1$ utilisé au sein de (2.33). Dans ce scénario, 10 sources sont exploitées, chacune émettant son information auditive de manière indépendante des autres : nous sommes donc dans



(A) Nombre de mouvements ($K_q = 0.1$)

(B) Nombre de mouvements ($K_q = 0.7$)

(C) Variabilité du comportement selon K_q

FIGURE 2.19 – Évaluation du nombre de rotation de la tête générées par le module MFI et le robot naïf. Les deux diagrammes polaires de gauche représentent les directions des 10 sources simulées, et la hauteur des flèches le nombre de mouvements générés (bleu : MFI, rouge : robot naïf), pour $K_q = 0.1$ (gauche) et $K_q = 0.7$ (droite). La sous figure de droite représente les mêmes données sous la forme d'histogrammes, pour 4 valeurs de K_q. Figure tirée de (COHEN-LHYVER, 2017).

un cas multisource. La figure 2.19 montre clairement l'effet de modulation du mouvement par le seuil K_q : lorsque sa valeur est faible, le système fait très confiance à son inférence, et peu de mouvements de tête sont nécessaires pour supposer connaître le lien entre les classes audio et visuelle. Au contraire, augmenter sa valeur conduit à une augmentation du nombre de mouvements de tête, de sorte que le système chercher à confirmer les inférences effectuées par le M-SOM en allant vérifier activement la donnée manquante. Dans tous les cas, le nombre de mouvements de la tête reste sensiblement inférieur à celui obtenu depuis un robot naïf. Enfin, l'effet du seuil K_q permet d'envisager une modification dynamique du comportement exploratoire du robot en fonction de la tâche à réaliser. Une telle modulation du comportement reste en dehors des objectifs du module HTM, mais pourrait être réalisée par un ensemble de règles de comportement données a priori au robot à un module décisionnel de plus haut niveau.

Cette rapide évaluation du module MFI a permis de mettre en évidence ses performances en terme de correction des erreurs de classification des experts audio et/ou visuel, d'inférence de données manquantes, et de génération des ordres moteurs selon la qualité estimée de l'apprentissage. Disposant maintenant (i) d'un système capable de juger de l'importance à porter à un objet présent dans l'environnement (le module DW) et (ii) d'un système à même d'estimer les classes audio-visuelle de ces objets, nous pouvons maintenant les utiliser de manière conjointe. C'est cette évaluation conjointe des deux modules qui est proposée dans la suite.

Publication

La contribution présentée dans cette sous-section, détaillant la formalisation du module d'inférence et fusion multimodale, a donné lieu à l'article (COHEN-LHYVER, ARGENTIERI et GAS, 2016) publié et présenté au sein de la conférence ICA.

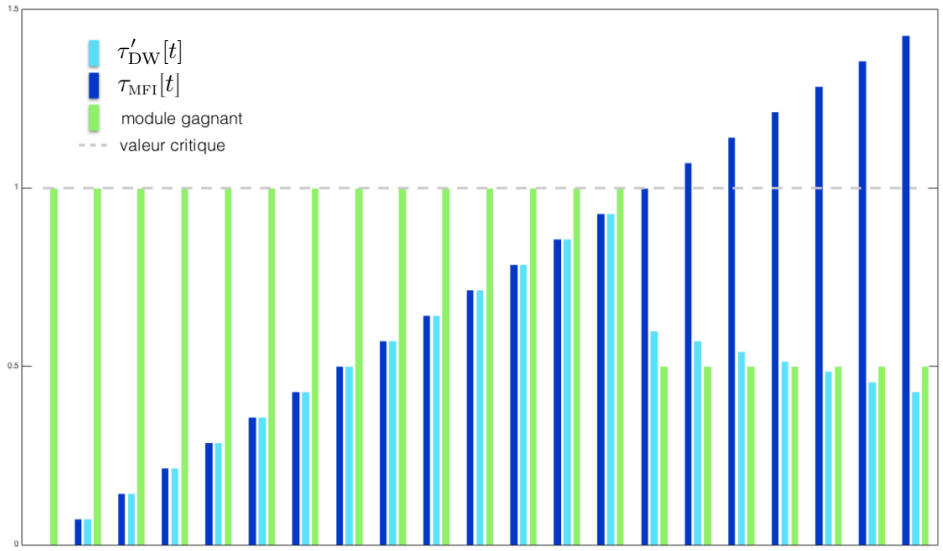


FIGURE 2.20 – Activité combinée (vert) des modules DW (bleu clair) et MFI (bleu foncé) en fonction du temps. Après une phase où l'inférence est prioritaire, la décision motrice revient au module DW, sur la base de la congruence. Figure tirée de (COHEN-LHYVER, 2017).

2.4.2.3 Évaluation conjointe des deux modules

En simulation : Chacun des deux modules constituant le module HTM est capable de générer un ordre moteur. Le module DW décide d'un mouvement de la tête selon une réaction attentionnelle, tandis que le module MFI en fait de même en réaction à une mauvais connaissance de l'environnement. Il est donc nécessaire de fusionner ces ordres moteurs, et par là même de décider d'une priorité entre ces deux modules. Dans la mesure où une décision motrice motivée par l'incongruence d'un objet ne peut se décider que sur la base d'informations audio et visuelle fiables, il semble naturel de privilégier le module MFI. Il en résulte ainsi une nouvelle expression de l'activité τ'_{DW} du module DW défini en (2.25) selon

$$\tau'_{DW}[t] = \tau_{MFI}[t] - \delta\left(\tau_{MFI}[t]\right) \times \tau_{DW}[t], \text{ avec } \delta(x) = \begin{cases} 1 \text{ si } x \geq 1 \\ 0 \text{ sinon.} \end{cases} \tag{2.35}$$

Nous voyons alors apparaître un comportement en 2 temps, illustré à la figure 2.20. Dans un premier temps, l'activité τ_{MFI} est inférieure à 1, signe que le système ne fait pas confiance en son inférence. A ce stade, le MFI est donc prioritaire (activité combinée égale à 1 sur la figure 2.20), et nous pouvons alors voir τ_{MFI} augmenter au fur et à mesure que la confiance en l'inférence augmente. Le MFI fait ensuite confiance à son inférence lorsque τ_{MFI} vaut 1. Alors selon (2.35) c'est τ'_{DW} qui devient plus petit que τ_{MFI} : c'est maintenant le module DW qui pilote les mouvements de la tête (activité combinée égale à 0.5 au sein de la figure 2.20). En pratique, ce comportement s'applique pour chacune des classes audiovisuelles des objets présents dans l'environnement, de sorte que certains objets bien appris sont gérés par le module DW, tandis que d'autres sont encore en cours d'apprentissage par le MFI. Au final, la congruence des objets n'est calculée qu'une fois que le système fait confiance en son inférence : le module MFI est donc comme indiqué précédemment prioritaire dans la décision motrice. Ce comportement en 2 temps est également illustré sur la figure 2.21 dans un cas unisource, où figurent dans le temps les objets focalisés par le robot (bas), ainsi que le module à l'origine de comportement moteur (haut). Après une phase où seul le MFI est en charge du comportement moteur, nous pouvons voir que l'objet de classe audio visuelle

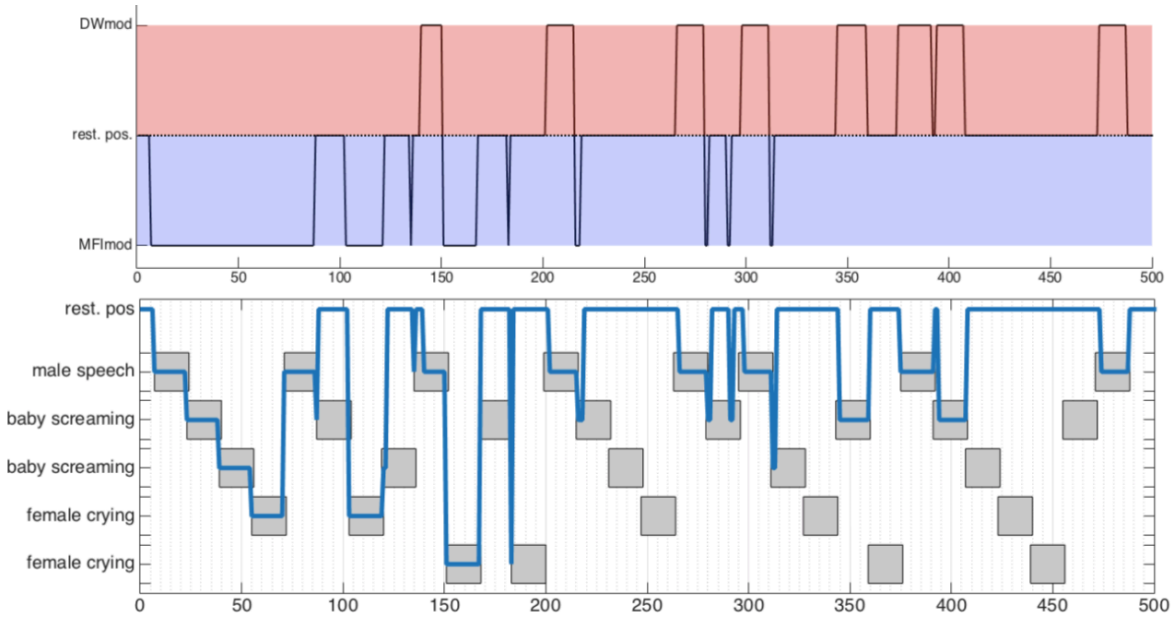
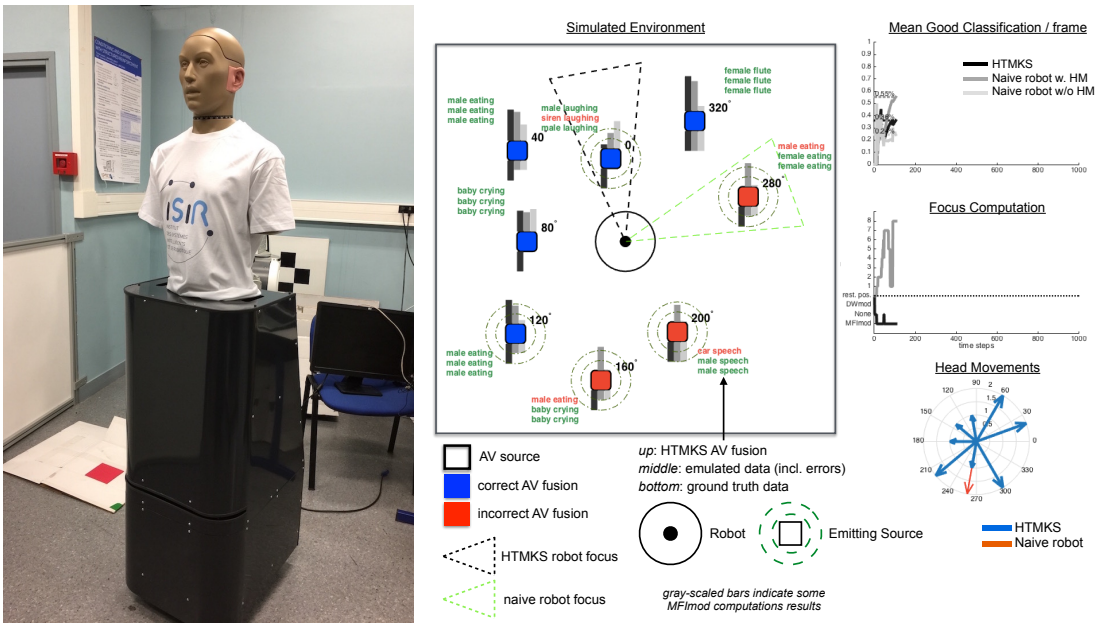


FIGURE 2.21 – Dynamique des deux modules DW et MFI, utilisés simultanément au sein du module HTM. Activité motrice déclenchée par le MFI ou DW en fonction du temps (haut), et objet focalisé en conséquence (bas). Figure tirée de (COHEN-LHYVER, 2017).

(speech,MALE) est bien appris, mais reste focalisé de par son incongruence. Le même comportement se produit plus tard pour les autres objets de classe audiovisuelle différente, mais au final c'est bien le module DW qui pilote le comportement moteur du robot. Ainsi sur la fin de la simulation, c'est bien uniquement l'objet (speech,MALE) qui est systématiquement focalisé car jugé incongru compte tenu de la probabilité d'occurrence de sa catégorie. Toute occurrence des autres objets est correctement inférée, et le module DW peut bien calculer les congruences de ces objets, même sans avoir accès à toutes leurs données audio et/ou visuelle.

D'autres évaluations concernant toutes ou partie des tâches propres aux 2 modules DW ou MFI ont été également évaluées à l'occasion de leur utilisation conjointe. Que ce soit le calcul de congruence, le nombre total de mouvement de tête généré, ou les taux de bonne reconnaissance des classes audiovisuelles des objets présents autour du robot, tous les résultats montrent un comportement analogue à celui obtenu lors des études isolées (des modules seuls) synthétisées dans les sous- parties précédentes. Néanmoins, ces évaluations ont été conduites en simulation, sur la base de sorties de classifieurs simulés. Si cette approche permet d'évaluer les différents modules dans des contextes expérimentaux difficiles à mettre au point en pratique de manière quantifiable et répétitive, il reste néanmoins à implémenter l'ensemble de l'architecture au sein du système TWO!EARS, et sur la plateforme mobile du projet.

Sur le robot réel : L'architecture complète du module HTM a été implémentée au sein de l'architecture du projet TWO!EARS sous la forme d'une "Knowledge Source", au même titre que tous les experts de localisation ou de reconnaissance. Il serait bien trop long de détailler cette implémentation, aussi le principal élément à retenir est que dans la suite ce sont les véritables systèmes de localisation et de reconnaissance binaurale proposés par les partenaires du projet qui sont utilisés. La localisation s'appuie en particulier sur un réseau de neurone profond, tandis que les différents experts de reconnaissance audio exploitent un système à base de GMM. Les détails théoriques et d'implémentation de ces différents algorithmes sont disponibles dans (KOLOSSA et BROWN, 2016). En particulier, les données



(A) Le robot ODI (B) Représentation interne

FIGURE 2.22 – Expérimentations : (gauche) robot ODI de l'ISIR, utilisé à l'occasion du projet TWO!EARS, (droite) aperçu de la représentation interne, utilisée en simulation et sur le robot, indiquant en temps réel les objets perçus par le robot ainsi que leur classe audio et visuelle.

utilisées pour constituer les différentes bases de données nécessaires à toutes ces approches ont été constituées à l'occasion de (longues) campagnes de mesure effectuées au sein de la salle d'expérimentation au LAAS-CNRS, réplique d'un appartement d'une centaine de mètres carrés sans plafond. En comparaison, une partie des évaluations proposées s'appuie sur des manipulations effectuées à l'ISIR, dans des conditions acoustiques très différentes. En conséquence, et comme déjà discuté et caractérisé au sein de §2.2.3.2, nous serons dans une situation où les conditions d'apprentissage et d'utilisation seront (très) différentes. Nous pouvons donc nous attendre, dans ces conditions, à des erreurs de classification et de localisation importantes, néanmoins gérées par le module MFI.

Les évaluations ont été conduites sur le robot ODI (voir figure 2.22a). Cette plateforme est constituée d'une plateforme mobile non holonome, sur laquelle est placée un simulateur de tête et torse (HATS) KEMAR. Sa tête est motorisée, via l'ajout d'un moteur pilotable depuis le robot et situé dans le torse. Une caméra est placée au dessus de la tête pour fournir les informations visuelles[7] extraites depuis des QR codes pour faciliter le traitement des informations visuelles (cette simplification n'a pas été effectuée sur le robot du LAAS, sur lequel le module HTM a également été utilisé). Les expérimentations conduisent à toute ou partie de la figure 2.22b, où figure la représentation interne de l'environnement construite au fur et à mesure de l'exploration du robot, et indiquant les mouvement de la tête et leur origine (DW ou MFI), ainsi que les éventuelles erreurs (corrigées ou non) de classification (pour les simulations uniquement). Une première évaluation consiste à vérifier les capacités d'évaluation des classes audiovisuelle du système HTM sur la base de données réelles. Dans le cadre d'un scénario impliquant 5 objets audiovisuels placés autour du robot, le taux de bonne estimation des classes de ces objets est représenté à la figure 2.23a au fur et à mesure de l'expérimentation. On peut y voir un taux de bonne classification convergeant vers environ 70%, contre seulement 38% pour le robot naïf s'appuyant uniquement sur les sorties

7. Le robot JIDO au LAAS est lui équipé de lunettes fournissant une vision stéréoscopique.

des classifieurs. Ces résultats expérimentaux sont tout à fait en adéquation avec les simulations reproduisant le même type de conditions, témoignant ainsi de la faculté du système à consolider via l'apprentissage du lien entre les modalités audio et visuelle sa représentation de l'environnement. Mais l'impact du module MFI ne se limite pas à la simple correction des probables erreurs de classification des experts disponibles. Tirant partie de l'expérience passée du robot consolidée dans le temps, le système complet réduit de manière significative l'espace des possibles sur lequel le module DW va devoir se baser pour conclure quant à la congruence des objets, comme indiqué sur la figure 2.23b. Là où le robot naïf doit composer avec possiblement 22 appariements audio et visuels possibles (la plupart étant produits par des erreurs de classification), le module HTM n'en propose plus que 5, dont 4 effectivement présents dans l'environnement. Il reste néanmoins encore une erreur de fusion, née de la combinaison entre les classes audio et visuelle de deux objets perçus au même moment (le cas multisource n'étant pas encore parfaitement géré au sein de l'architecture TWO!EARS des experts).

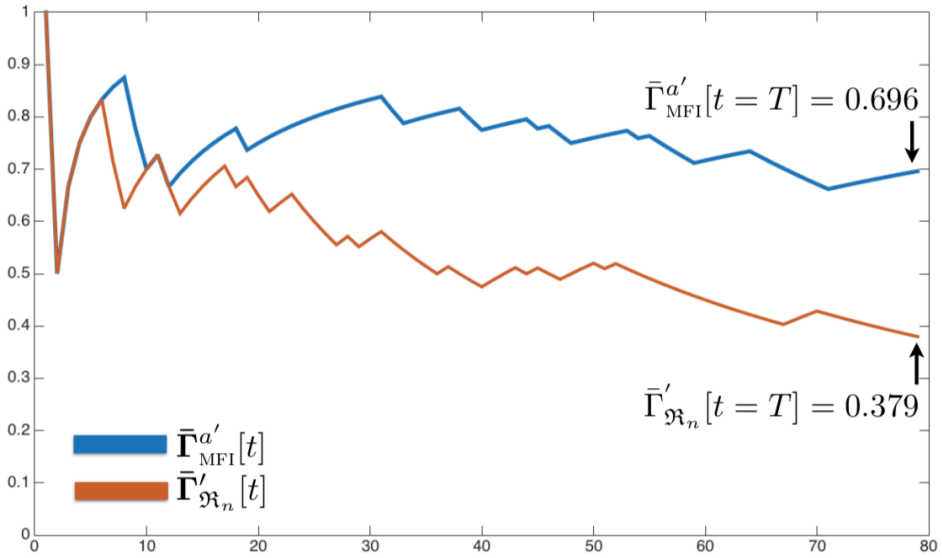
Publication

Le système HTM en général, ainsi qu'une partie de l'évaluation synthétisée dans cette sous-section, a donné lieu à l'article (COHEN-LYVER, ARGENTIER et GAS, 2018) publié dans la revue "Frontiers in Neurorobotics" au sein du *Research Topic* intitulé "Intrinsically Motivated Open-Ended Learning in Autonomous Robots".

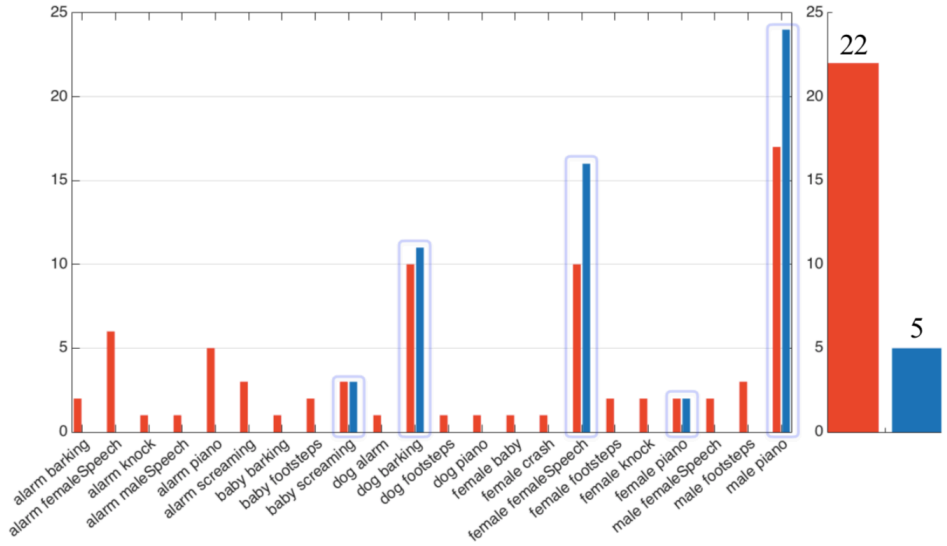
2.4.3 Architecture logicielle pour l'audition binaurale

Alors que l'essentiel des contributions précédentes étaient méthodologiques, nous souhaitons pour terminer mentionner une contribution pratique importante effectuée à l'occasion du projet TWO!EARS. Antonyo MUSABINI a en effet travaillé en tant qu'ingénieur sur ce projet avec pour objectif l'implémentation logicielle complète de la chaîne d'acquisition audio et de ses traitement au sein du middleware ROS. Son travail d'ingénierie est rapidement évoqué dans la suite.

L'architecture logicielle du projet TWO!EARSs'appuie sur une première étape de traitements des signaux binauraux, regroupés au sein d'un *Auditory Front-End* (AFE) en charge d'effectuer la séparation en bande de fréquences via des filtres gammatone, de reproduire la transduction des cellules ciliées de la cochlée, puis d'extraire des caractéristiques binaurales (corrélations, indices interauraux, etc.). Cet AFE a d'abord été codé en Matlab par les partenaires experts des modèles d'audition binaurale sous Matlab. Il est cependant difficile d'envisager, comme souvent, son utilisation dans un contexte robotique quasi temps-réel. Par ailleurs, à notre connaissance, très peu de systèmes logiciels bas-niveau *dédiés à l'audition binaurale* sont disponibles dans la communauté. Nous avons donc travaillé en collaboration avec le LAAS-CNRS à la définition d'un AFE binaural sous ROS permettant de calculer en temps réel les caractéristiques audio nécessaires au projet. Pour cela, Antonyo MUSABINI a codé sous notre supervision un module ROS nommé rosAFE permettant l'assemblage dynamique de briques de traitements élémentaires –appelées processeurs– toutes spécifiées sous la forme d'une machine d'état standard. rosAFE permet alors de formaliser leurs connexions les uns aux autres (série ou parallèle, cf. figure 2.24a) sous la forme de réseaux de Pétri, comme représenté aux figures 2.24b et 2.24c. L'architecture proposée garantie la bonne transmission des données entre chaque processeur au plus rapide de leurs cadences computationnelles respectives : dès qu'une donnée issue d'un processeur est disponible, elle peut potentiellement être transmise à tous ceux en dépendant. Les temps de



(A) Taux de bonne classification audiovisuelle



(B) Classes audiovisuelles créées

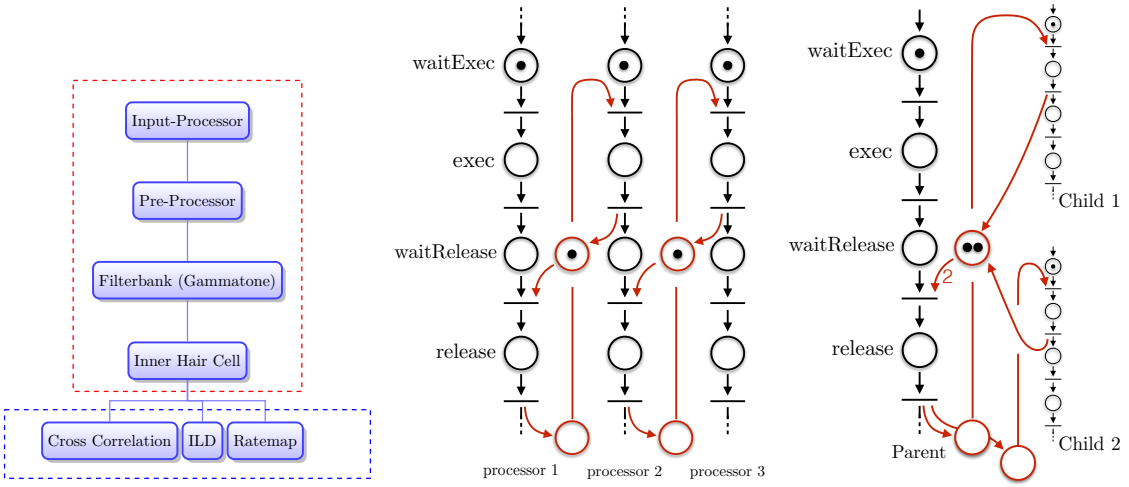
FIGURE 2.23 – Résultats expérimentaux. (Haut) Taux de bonne classification audiovisuelle, pour le module HTM au complet (bleu), et comparé au robot naïf (rouge), au cours du temps. (Bas) Classes audiovisuelles rencontrées au cours de l'expérimentation. Le robot naïf en rencontre 22, résultat des associations multiples des différentes sorties (erronées) des classifieurs. Le système HTM n'en propose plus que 5 (entourées), pour 4 effectivement présentes. Figure tirée de (COHEN-LHYVER, 2017).

Processeur	Matlab	openAFE	Ratio
Input + normalisation	1e-4s	7.3e-6s	13.7
Pre-Proc (DC filter)	1.1e-3s	7.8e-6s	141.5
Pre-Proc (Pre emphasis)	1.2e-3s	9.09e-6s	132
Pre-Proc (RMS normalisation)	1.2e-3s	3.05e-5s	39.4
Pre-Proc (Level scalling)	1.6e-3s	9.08e-6s	176.2
Gammatone	5.1e-3s	0.968e-3s	5.3
IHC	2.8e-3s	0.73e-3s	3.84
ILD	4.3e-3s	1.16e-3s	3.71
Ratemap	6.8e-3s	1.83e-3s	3.71
Cross-correlation	26.2e-3s	30.6e-3s	0.86

TABLE 2.2 – Temps de calcul moyen sous Matlab ou ROS pour chacun des processeurs.

calcul comparés entre l'implémentation Matlab et ROS de l'AFE figurent dans le tableau 2.2. Mis à part pour le processeur dédié au calcul de l'intercorrélation [8], rosAFE permet d'aller en moyenne 50 fois plus vite que l'AFE sous Matlab. Par ailleurs, il est possible d'ajouter, supprimer, modifier, au cours de l'exécution tout ou partie des processeurs. Cette flexibilité est selon nous inédite et permet de modifier à la volée la plupart des paramètres des différents algorithmes utilisés pour l'extraction de caractéristiques audio. Au final, les données calculées par cet AFE peuvent être exploitées par n'importe quel autre module ROS, ou alors transmises à un client Matlab via un bridge avec ROS. L'intégralité de ces développements ont été rendus publics [9], sont documentés [10], et disponible sur un serveur GIT [11].

8. Ses mauvaises performances sont très probablement liées à un bug d'implémentation.
9. http://www.twoears.eu
10. http://docs.twoears.eu/en/1.4/robo/rosafe/
11. https://github.com/TWOEARS/rosAFE



(A) Arbre de processeurs. (B) Concurrence verticale. (C) Concurrence horizontale.

FIGURE 2.24 – Illustration du fonctionnement du module rosAFE. (A) Succession de processeurs, de manière sérielle (concurrence verticale) ou parallèle (concurrence horizontale). (B) & (C) Réseaux de Pétri pour les deux types d'organisation. A tout moment, les processeurs traitent les données qui leur sont accessibles : le réseau de Pétri proposé permet de gérer la concurrence à l'accès aux données et garantie qu'un processeur ne peut récupérer et traiter des données que quand il en a le droit.

2.4.4 Conclusion

Comment l'action peut-elle aider à l'analyse de la scène sonore ? Dans les travaux présentés dans cette sous section, nous en avons eu une utilisation inédite : l'action permet d'aller chercher les données multimodales manquantes, et éventuellement de renforcer les connaissances apprises sur l'environnement. L'architecture proposée a été utilisée expérimentalement au sein du démonstrateur final du projet TWO!EARS, démontrant ainsi l'apport du comportement actif du robot au sein d'une architecture finalement classique d'analyse du flux audio. Néanmoins, certains aspects dans cette analyse n'ont pas été totalement traités. Par exemple, une des principales limites du module DW réside dans la non prise en compte de l'aspect temporel au sein de la définition de la Congruence : en l'état actuel de la formalisation, aucune habituation à un quelconque stimulus n'est capturée. Par ailleurs, l'importance prise par les experts de localisation est vraisemblablement trop grande : si l'expert dédié à la localisation ne fournit pas une bonne estimée de l'azimut de la (ou les) source(s), le robot pourrait se retrouver à faire face à un autre objet que celui émettant effectivement le son. Pour résoudre ce problème, une piste intéressante consisterait à inclure au sein du MSOM l'estimation de la position de la source : le MFI pourrait alors gérer conjointement les erreurs de reconnaissance et de localisation. Enfin, alors que le DW a l'ambition d'être capable de traiter tout type de scène sonore sans trop d'a priori, il nécessite néanmoins un environnement statique, aucune stratégie de suivi (spatial et temporel) n'ayant été incluse au système. Pour terminer, on notera la généricité de la définition d'objets proposées : n'importe quel type de caractéristique peut y être incorporée en plus de la simple classe audio ou visuelle. Il suffit de disposer pour cela des experts dédiés à ces nouvelles caractéristiques (par exemple, la reconnaissance d'émotions, de hauteur d'un son, etc.) pour compléter de manière pertinente les informations fusionnées au sein du MFI. La représentation de l'environnement n'en serait alors que plus riche.

Chapitre 3

Contributions pour une approche interactive de la perception

Sommaire

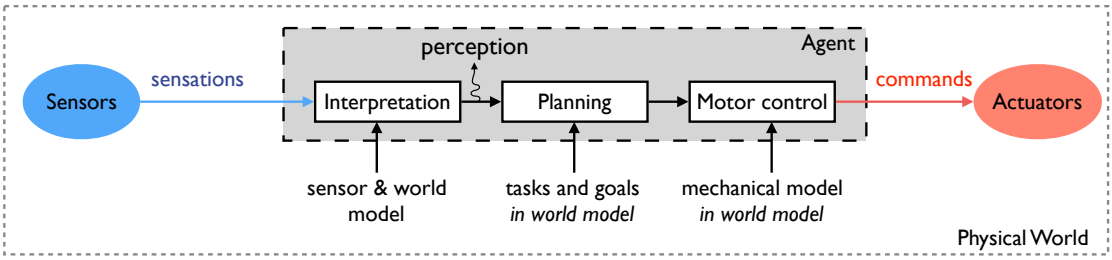
Les travaux présentés dans la Section 2, dédiés à la modalité auditive, ont permis de montrer en quoi le mouvement était susceptible d'améliorer la perception qu'un robot pouvait avoir de l'environnement. Au delà des conséquences triviales de l'action (i.e. un déplacement du robot à prendre en compte vis à vis des changement acoustiques résultants), l'ensemble des contributions précédentes mettent en place une chaîne de traitement de l'information dont l'objectif est l'analyse *objective* de la scène sonore. Nous sommes donc face à

une approche computationnelle qui d'abord analyse les informations provenant des microphones (extraction de caractéristiques), en interprète le contenu via des modèles donnés par l'ingénieur (relation spatiale entre les caractéristiques et la position des sources), raisonne sur ce résultat pour planifier l'action dans le but de réaliser une tâche bien identifiée (exploration, reconnaissance de sources, etc.), et génère en conséquence des actions motrices. Cette organisation de type *sentir-planifier- agir*, très classique aujourd'hui en robotique, a originalement été proposée dans les années 1950 lors de l'émergence de problématiques d'Intelligence Artificielle au sein d'agents artificiels incarnés (MARR, 1982 ; RUSSELL et NORVIG, 2003). Utilisée largement par d'autres modalités perceptives (et en particulier visuelle), cette approche permet encore aujourd'hui de réaliser des robots capables de réaliser des tâches complexes (saisie d'objets, déplacement autonome dans des environnements inconnus, etc.) de manière efficace ; nous l'avons en particulier illustré dans le cadre d'une tâche audio d'exploration, cf. 2.4. Nous allons illustrer dans ce nouveau chapitre en quoi cette approche classique de la perception (qu'elle soit sonore ou non) introduit de nombreux problèmes et connaît de nombreuses limites. Ces points seront abordés dans une première section d'introduction, précisant en particulier les enjeux d'une nouvelle approche que nous appellerons "interactive" de la perception. Sur cette base, nous proposons de présenter dans les sections suivantes nos différentes contributions sur le sujet : tout d'abord sur l'estimation de la dimension de l'espace (section 3.2), puis concernant l'extraction d'une représentation des invariants sensorimoteurs (section 3.3).

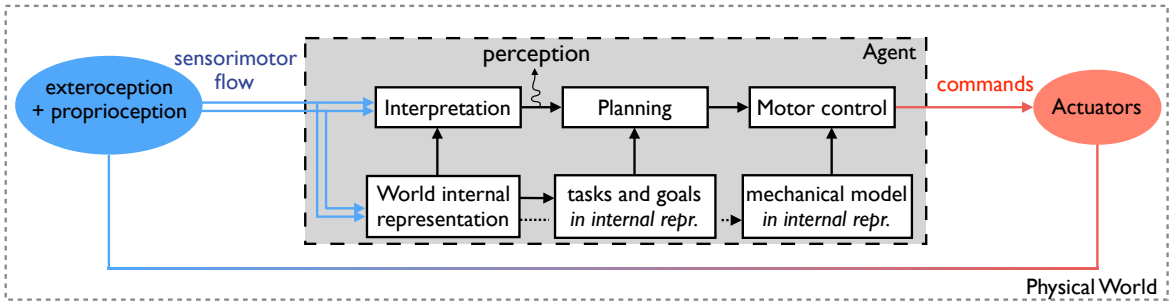
3.1 Introduction

3.1.1 Contexte

Comme nous avons eu l'occasion de l'illustrer dans le chapitre précédent, la plupart des méthodes d'analyse de la scène sonore ont ceci de commun qu'elles s'appuient fortement sur la connaissance de modèles a priori, fournis par l'ingénieur à l'occasion de la programmation des méthodologies. Par exemple, les contributions détaillées en §2.2 s'appuient sur la physique de la propagation (champ libre, modèles de la tête, etc.) pour modéliser, caractériser et identifier le lien existant entre les caractéristiques des signaux binauraux (indices binauraux, monauraux, etc.) et l'origine spatiale des sources sonores à leur origine (azimut, élévation, distance). De manière moins directe, mais toute aussi explicite, les contributions proposées en §2.4 s'appuient également sur de telles connaissances a priori : étape de localisation déjà réalisée, modèle géométrique du robot permettant de savoir où tourner la tête en fonction des résultats de localisation, etc. Cette remarque reste bien sûr valide pour beaucoup d'autres approches visant à exploiter d'autres modalités : par exemple, reconnaître les expressions sur un visage nécessite de disposer d'un modèle de celui-ci, fourni soit de manière analytique, ou par apprentissage préalable (DAPOGNY, BAILLY et DUBUISSON, 2018). Pour autant, peut-on dire que le robot *perçoit* son environnement ? L'ensemble des travaux traitant de telles problématiques se réclament du domaine de la perception robotique, entendue ici comme étant la faculté à interpréter les données issues des capteurs équipant le robot. Mais cette faculté d'interprétation est ici restreinte au contexte courant dans lequel il est prévu que le robot opère, pour un but préalablement fixé par le programmeur, tel qu'illustré à la figure 3.1a. La perception robotique traditionnelle s'implémente ainsi au sein d'une boucle sensation/planification/action, s'appuyant sur les connaissances a priori du monde, de la structure de l'agent robotique et de son interaction avec son environnement pour faire émerger une interprétation des sensations (comprises ici comme étant les données brutes issues des capteurs). Comme déjà illustré à la figure 2.2 du chapitre précédent, cette architecture a donné lieu au champ de la *perception active* (CHAUMETTE, 1998), dont



(A) Boucle de perception classique



(B) Vision sensorimotrice de la perception

FIGURE 3.1 – Comparaison de l'approche "perception/planification/action" traditionnelle, impliquant la boucle classique de perception telle que déjà illustrée à la figure 2.2, avec la version "sensorimotrice", dans laquelle l'action est indissociable de la perception.

les premières bases sont posées au début des années 90 (ALOIMONOS, WEISS et BANDOPADHAY, 1988 ; BAJCSY, 1988). Il s'agit alors de palier, via la mise en mouvement du capteur, à l'insuffisance des modèles face aux mesures expérimentales bruitées, au manque de données, aux ambiguïtés : l'action est alors pensée comme un outils permettant d'améliorer la perception. Néanmoins, la plupart des approches actives de la perception restent peu génériques et doivent toujours être adaptées au contexte applicatif ciblé. Pour faire face à cette difficulté, des modèles probabilistes ont été introduits, comme les modèle bayésiens de la perception (KNILL et RICHARDS, 1996 ; BESSIERE, LAUGIER et SIEGWART, 2008) cherchant à reconstruire l'interprétation la plus vraisemblable d'entrées sensorielles possiblement ambiguës en termes objectifs, fournis par le programmeur. On comprend donc qu'en perception robotique, contrairement aux systèmes vivants, ce n'est par l'agent qui interprète les données, mais bien le programmeur, via l'algorithme qu'il a conçu et placé au sein du robot. Dès lors, pour espérer pouvoir réaliser une tâche donnée, un robot doit être préalablement doté *par le programmeur* de tous les modèles, mais également de tous les concepts, dont il pourrait avoir besoin pour la réaliser. Cela inclue donc à la fois les aspects bas niveaux (basés signaux), mais également haut niveaux (logiques, symboliques, pour la décision et la cognition), qui se doivent alors de couvrir tous les cas possibles, classiquement rencontrés dans des environnements ouverts et inconnus. Autant dire que cela n'est pas envisageable.

Alors, comment envisager que tout ou partie de ces aspects puissent être découverts, construits, appris, sans recours à une quelconque intervention extérieure ? Cette question intéresse la Robotique Autonome, dont l'objectif est justement de doter les robots de capacité d'autonomie et d'adaptation (KHAMASSI et DONCIEUX, 2016). Les capacités perceptives des robots se trouvent être au cœur de cette problématique, car à l'interface entre les sensations et les étages cognitifs du robot. Pourtant, on a longtemps cru que de telles capacités d'autonomie pouvaient être atteintes uniquement via l'implémentation symbolique de raisonnements, prenant le parti d'oublier peut être l'incarnation de cette autonomie au sein d'un agent robotique en interaction permanente avec son environnement. Cela ne réduit en

rien la portée des résultats obtenus avec cet type d'intelligence : dans des contextes bien identifiés et limités, et aidés par des capacités de calcul énormes, des programmes informatiques peuvent aujourd'hui reconnaître des objets au sein d'images naturelles, reconnaître le contenu d'un signal de parole au sein d'une scène sonore très bruitée, ou même battre les humains aux échecs ... mais sans pour autant avoir la capacité d'en déplacer les pièces sur n'importe quel échiquier. A nouveau se pose donc la question de savoir si l'agent robotique perçoit son environnement via ce type de paradigme. A ce stade, la perception reste reléguée à une interprétation des sensations, alors que l'incarnation d'un agent robotique en interaction réciproque et permanente avec son environnement via son propre corps conditionne les connaissances qu'il est susceptible d'en extraire. Il apparaît alors difficile de faire abstraction du rôle de l'action, i.e. de la façon donc l'agent explore et interagit activement avec l'environnement, dans une description qui se voudrait complète. Et alors qu'on pensait la perception comme une notion instantanée, sorte de photo des champs physiques interagissant avec les capteurs, il semble dans un tel paradigme que la perception ne puisse être envisagée –voir même seulement exister– sans l'action. Dès lors, le flux sensorimoteur (et non plus seulement les sensations) serait le porteur des informations sur l'interaction de l'agent avec son environnement. Plusieurs expériences en psychologie de la perception plaident pour cette vision active de la perception (BERNARD, 2014). Ainsi, les conséquences de la privation de mouvement sur la perception ont été illustrées dans (HELD et HEIN, 1963) à l'aide d'une expérience impliquant deux chats nouveaux nés. Placés tous deux dans un environnement constitué de bandes noires et blanches verticales, un premier chaton libre de bouger provoque le déplacement du second chaton attaché dans un chariot. Testés sur des tâches de *reaching* et d'équilibre, seuls les chatons passifs, pour lesquels les variations sensorielles n'étaient pas liées à une quelconque action voulue de leur part, ont montré un comportement inadapté. Cette première expérience suggère que le flux sensoriel seul n'est pas suffisant lors de la phase d'exploration de l'environnement pour développer un comportement adapté. Un autre argument peut être tiré des phénomènes de substitution sensorielle. Concept introduit à la fin des années 60 (BACH-Y-RITA et al., 1969 ; BACH-Y-RITA et KERCEL, 2003), il s'agit de substituer une modalité sensorielle par une autre via un dispositif dédié. Dans cet exemple, il s'agit de remplacer la modalité visuelle (ici représentée par une caméra placée sur des lunettes) par la modalité tactile : l'image de la caméra est retranscrite sous la forme d'impulsions mécaniques produites par une matrice d'aiguilles placées sur le ventre ou le dos. Après une phase d'apprentissage, le patient est capable d'identifier et localiser des objets, l'expérience tactile s'effaçant au profit d'une sensation d'externalisation des stimuli. Mais là encore, ceci n'est possible que si la caméra est mise en mouvement par le patient lui même lors de la phase d'apprentissage. Sans exploration active, les sujets ne ressentent que des sensations tactiles difficiles à interpréter. D'autres exemples peuvent être listés, comme (RICHTERS et ESKEW, 2009) où il est montré que la perception des couleurs peut être modifiée en changeant les caractéristiques de l'interaction sensorimotrice avec l'environnement, ou encore (PHILIPONA et KEVIN O'REGAN, 2006) cherchant à expliquer notre perception des couleurs "pures" par les propriétés de nos interactions sensorimotrices avec des surfaces colorées.

La figure 3.1b tente de synthétiser ce nouveau paradigme, dit sensorimoteur, de la perception. Dans ce schéma, la perception n'est plus simplement une interprétation des sensations sur la base d'une représentation interne fournie *a priori*, mais plutôt une interprétation d'une représentation du monde construite sur la base du flux sensorimoteur au cours de l'interaction de l'agent robotique avec son environnement. Cette représentation, dont nous discuterons plus tard la forme, n'est plus fournie a priori, mais modelée, mais aussi limitée, par les capacités d'action et de sensation du robot : d'une représentation objective, obtenue depuis un point de vue externe à l'agent, nous passons à une représentation subjective (BRETTE, 2013), interne et propre au robot. Selon ce point de vue, percevoir n'est donc

plus une capacité fournie en amont à l'agent, mais une faculté découverte et apprise progressivement de manière active (MAILLARD et al., 2005). La théorie des contingences sensorimotrices (SMC) proposée au début des années 2000 par K. O'Regan et A. Noé J. K. O'REGAN et NOË, 2001 aborde la question de la perception selon ce paradigme sensorimoteur, mais principalement sous un angle philosophique, abordant notamment le problème de la conscience dépassant le cadre de ce manuscrit. On en retrouve les prémices dans l'approche écologique de la perception selon Gibson (GIBSON, 1979) pour qui il faut considérer l'agent dans sa relation avec son environnement pour comprendre comment percevoir le monde. D'un point de vue plus formel, on en retrouve également la trace dans les réflexions de Poincaré concernant les fondements de la géométrie et de l'espace (POINCARÉ, 1887). L'intuition de Poincaré est que notre perception de l'espace n'est pas une faculté innée, mais qu'elle se construit via notre expérience sensorimotrice, en mettant en relation les retours perceptifs de nos propres actions. Et Poincaré suggère alors que certaines structures au sein du flux sensorimoteur sont susceptibles de capturer des propriété, parfois spatiales, de l'interaction avec notre environnement. L'intérêt des réflexions de Poincaré réside dans la remise en perspective de la notion d'espace, pourtant a priori très naturelle. A vrai dire si naturelle qu'il est rare que cette notion ne soit pas un prérequis évident à l'ensemble des travaux cités précédemment. Définir un référentiel au sein d'un espace forcément supposé Euclidien est pour le physicien, le roboticien, une des tâche les plus commune préalablement à la définition d'une tâche s'exprimant selon ce référentiel. Pourtant, si nous adoptons le point de vue de l'agent, qui n'a donc accès qu'à son unique flux sensorimoteur, définir l'espace comme cette sorte de contenant commun de la matière n'est plus si évident. D'ailleurs, cette définition est même maintenant totalement contre-intuitive : là où le physicien décrit un espace homogène, isotrope, continu, et 3D, l'agent a accès à des sensations hétérogènes (provenant de capteurs de nature différente), discrètes, quantifiées, et vraisemblablement à très haute dimension. A nouveau, ce n'est pas tant ces propriétés des sensations qui permettent de ressentir l'espace, mais plutôt leurs caractéristiques invariantes –les contingences– de l'interaction avec l'environnement.

Publication

Ces éléments de contextualisation ont été en partie abordés dans l'article (GAS et ARGENTIERI, 2016) publié au sein de la revue francophone "Intellectica". Cet article propose une introduction à la perception sensorimotrice en robotique.

3.1.2 Positionnement et ligne de recherche

La théorie des contingences sensorimotrice semble donc proposer une solution particulièrement intéressante pour les problématique d'autonomie mentionnées précédemment. Disposer d'un agent robotique capable de construire de lui-même une image de son interaction, modulée par ses propres capacités d'action et de perception, permettrait de doter des robots de réelles capacités d'adaptation inenvisageable dans le paradigme classique. Si l'agent perd certaines facultés motrices (destruction d'un moteur), ou perceptives (caméra défaillante), il serait dès lors capable de redécouvrir ou d'adapter ses contingences à sa nouvelle condition, sans autre intervention externe. Cependant, dans sa présentation initiale, l'approche SMC souffre de ne pas être formellement décrite, ce qui peut conduire certains travaux plus proches du paradigme de la perception active "traditionnelle" de s'en réclamer. D'ailleurs, nous avons nous même longtemps utilisé la terminologie "perception active" pour désigner certains travaux pourtant en relation directe avec la théorie SMC. Pour bien distinguer ces deux approches intrinsèquement différentes de la perception, il semble donc pertinent d'introduire une nouvelle terminologie. Il a été proposé récemment dans (BOHG et

al., 2017) d'utiliser le terme de *perception interactive* pour désigner l'ensemble de approches "qui exploitent n'importe quel type d'interaction puissante avec l'environnement afin de simplifier et améliorer la perception". Nous aurions envie d'ajouter à cette définition que c'est même l'interaction elle même qui définit la perception au sens de la théorie SMC. Telle que définie dans (BOHG et al., 2017), l'approche SMC peut se voir comme la brique la plus bas niveau de la perception interactive, celle qui dans ce même papier est indirectement qualifiée de question ouverte.

C'est dans ce cadre que nous avons (i) cherché à vérifier l'applicabilité de l'approche SMC à la robotique, mais également (ii) travaillé à découvrir quels invariant de l'interaction agent/environnement pouvaient expliquer en quoi le flux sensorimoteur est porteur d'informations permettant de faire émerger une perception du monde. Ces deux points ont été abordés à l'occasion de la direction de 2 thèses effectuées par Alban LAFLAQUIÈRE et Valentin MARCEL dont les travaux sont synthétisés dans les deux sections à venir. Systématiquement, ces problématiques ont été abordées selon la ligne de recherche suivante :

— l'agent robotique (simplement dénommé *agent* dans la suite) sera supposé naïf, sans aucune connaissance a priori sur la structure physique de son corps, de ses capteurs, de ses actionneurs. Il ne dispose d'aucune indication sur les propriétés physiques de son interaction avec l'environnement. Sa seule donnée accessible est son flux sensorimoteur, constitué des commandes motrices et de sa perception ;

— on suppose également que les commandes motrices sont ressenties via la proprioception de l'agent, celle-ci étant alors définie comme une copie efférente des ordres moteurs. Cela revient à imaginer une relation bijective entre la commande, supposée instantanée ou à dynamique rapide, et la proprioception. De fait, les sensations accessibles à l'agent sont donc supposées déjà séparées en proprioception et extéroception. Si cette séparation semble naturelle selon les approches traditionnelles en robotique, elle pose la question de la découverte de cette catégorisation des sensations, peu abordée à notre connaissance, et pas aussi naturelle selon que l'on parle de robotique ou de systèmes vivants (GAPENNE, 2014) ;

— la naïveté du robot conduit naturellement à une difficulté à définir une tâche que le robot doit réaliser. S'il est toujours possible d'introduire des motivations particulières internes à l'agent (curiosité, survie, etc.), elles restent cependant définies par l'ingénieur a priori et sont donc à même de biaiser l'analyse du flux sensorimoteur. Une des conséquences immédiate est que pour la plupart des travaux présents dans la suite, l'exploration motrice effectuée par l'agent de son environnement sera aléatoire. Il est vraisemblable qu'une telle exploration soit sous-optimale, mais elle permet néanmoins de valider expérimentalement l'existence d'invariants au sein du flux sensorimoteur. Une fois ces invariants découverts et caractérisés, il sera alors possible de les exploiter pour une tâche donnée (et non, donc, l'inverse) ;

— nous tâcherons, autant que possible, d'identifier et de caractériser les hypothèses sous-jacentes à poser permettant l'analyse du flux sensorimoteur. Si certaines d'entre elles pourront paraître triviales dans un cadre robotique classique (comme la séparation proprio/extéroception mentionnée plus haut par exemple), il est important de comprendre leurs potentielles conséquences sur la caractérisation du flux sensorimoteur.

Quand se pose la question de comment capturer l'interaction d'un agent avec son environnement, la question de l'espace et de sa représentation par cet agent est probablement une des premières problématiques à traiter. De sa connaissance dépend directement la capacité de l'agent à se déplacer, à se situer, à situer des objets et plus tard à établir des schémas de navigation. Cette représentation permet à l'agent d'exprimer ses observations en termes de positions, de déplacements, du moins pour un observateur externe. Pour l'agent,

des termes comme "position", "distance", se doivent d'être définis relativement à des propriétés intrinsèques du flux sensorimoteur, propriétés qu'on espère partagées avec l'environnement : l'espace apparaît alors comme un contenant commun et partagé dans lequel gravite l'agent et son environnement (nous faisons d'ailleurs là l'hypothèse que les deux sont clairement séparés, ce qui semble non trivial). Il est clair que par ailleurs certaines caractéristiques peuvent être non partagées avec l'agent, du moins en être indépendant : on parlera alors de caractéristiques non spatiales, comme la couleur, la température d'un objet. La question de la perception de l'espace est une des questions centrales soulevées par Poincaré, et nous avons eu l'occasion de travailler à l'extension d'une première tentative de formalisation de ses idées, proposée initialement dans (PHILIPONA, J. K. O'REGAN et NADAL, 2003). Ces travaux visent à montrer sous quelles conditions un agent naïf est capable de découvrir la dimension de son interaction avec l'environnement, appelée par abus de langage "dimension de l'espace". Cette partie constitue une première contribution à l'approche SMC, effectuée à l'occasion de la thèse d'Alban LAFLAQUIÈRE (LAFLAQUIÈRE, 2013). Elle sera présentée dans la section 3.2 de ce manuscrit. Sur cette base, nous avons dans un premier temps proposé de caractériser les invariants sensorimoteurs, tout d'abord naïvement, sans formalisation mathématique poussée (cf. §3.3.1). Les résultats obtenus dans ce cadre laisse penser qu'une structure mathématique forte, et donc un formalisme, peut être exploité pour mieux comprendre et analyser ces invariants. C'est précisément l'objectif de la thèse de Valentin MARCEL que d'effectuer ce travail de formalisation. Sa thèse est toujours en cours, mais nous a permis d'ores et déjà de proposer une formalisation s'appuyant sur des considérations topologiques via l'introduction d'espaces quotient, appliquées à la découverte du corps de l'agent (§3.3.3) ou de son espace de travail (§3.3.4). L'ensemble de ces contributions sont synthétisées dans les sections suivantes.

3.2 Estimation de la dimension de l'espace

3.2.1 Poincaré et le groupe des mouvements compensables

3.2.1.1 Généralités

Comme précisé en introduction, il appartient à Poincaré d'avoir, peut être le premier, réfléchi au point de vue interne de la représentation de l'espace. Il nous rappelle en effet que nos espaces sensoriels possèdent des caractéristiques très éloignées de ce que serait l'espace géométrique, au sens d'un expérimentateur extérieur. Alors, comment faisons nous l'expérience partagée de cet espace géométrique sans y avoir accès par les sens (POINCARÉ, 1913) ? La réponse qu'il propose s'appuie sur la nature des caractéristiques spatiales partagées entre l'agent et son environnement. Si d'une part l'agent en question est capable de se déplacer, et que d'autre part l'environnement dans lequel se situe l'agent est constitué d'objets rigides, alors tous deux partagent des capacités de déplacements identiques. Et ce sont ces capacités de déplacement qu'on espère dotées de propriétés analogues à l'espace des géomètres : continuité, homogénéité, dimensionnalité, etc. Cette intuition prend forme via l'introduction de *transformations compensables*, qui se traduisent par un phénomène particulier dans l'expérience sensorimotrice de l'agent, cf. figure 3.2. Soit l'agent au repos, dans un état sensoriel donné. Si l'environnement subit une telle transformation compensable, alors l'état sensoriel de l'agent va changer. Cependant —et c'est là que l'on retrouve la nature "partagée" de l'espace— l'agent est capable via une action de retrouver sa sensation initiale : il a compensé certaines variations environnementales par l'action. Dans le cas contraire, il peut alors déduire qu'il est face à une transformation d'état de l'environnement qui est a priori non spatiale.

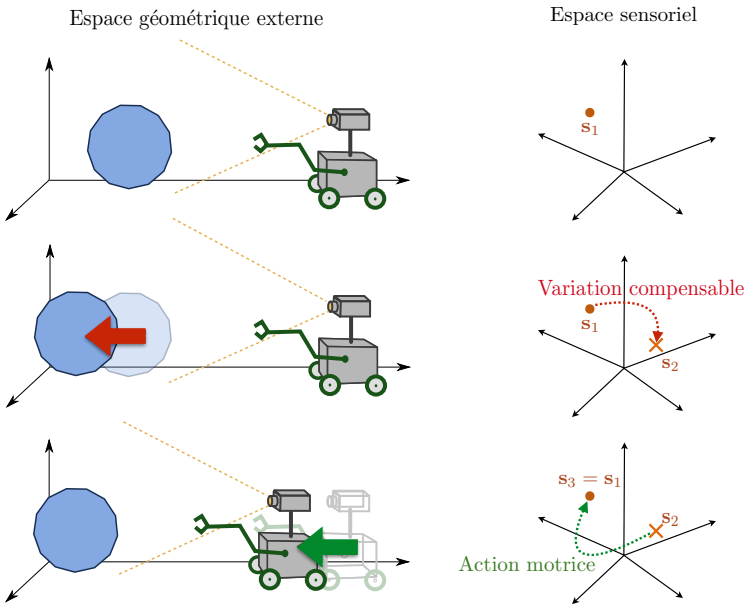


FIGURE 3.2 – Illustration de la notion de transformation compensable. Après une modification (compensable) de l'état de l'environnement, l'agent est capable de retrouver sa sensation initiale en se déplaçant dans l'environnement. Figure tirée de (GAS et ARGENTIERI, 2016).

Mais comment approcher ces transformations compensables et espérer alors capturer leurs propriétés ? Poincaré fait ici l'hypothèse que nous sommes capables, de manière innée, de former des groupes (au sens des structures algébriques de l'algèbre générale [1]) (POINCARÉ, 1902). Et il propose que la notion d'espace et ses propriétés peuvent se déduire de cette seule notion de groupe. Si l'hypothèse est osée, elle permet néanmoins d'imaginer qu'un agent ayant seulement accès à son flux sensorimoteur ait accès à la dimension de l'espace dans lequel il interagit avec son environnement. Pour revenir à un exemple robotique simple, nous pouvons illustrer ces idées grâce à un bras robotique à plusieurs degrés de liberté, fixe dans l'environnement, et équipé d'un capteur extéroceptif en son extrémité. Si le bras décrit des mouvements aléatoires, alors la variété proprioceptive (ou motrice, les deux étant selon les hypothèses précédentes identiques) est de dimension d_1 égale au nombre de degrés de liberté. Si maintenant l'extrémité du bras (et donc le capteur) est maintenue fixe dans l'espace, cette dimension est réduite du nombre de degrés de liberté dans l'espace d, de sorte que la variété proprioceptive possède maintenant une dimension $d_2 = d_1 - d$. Au final, il est donc possible d'avoir accès à la dimension d selon $d = d_1 - d_2$. Dans ce raisonnement, la dimension d est obtenue depuis la variété motrice (nous reviendrons d'ailleurs plus longuement sur cette notion de variété dans la suite), semblant réduire le rôle des sensations extéroceptives sur la compréhension de l'environnement. Pourtant, d'un point de vue interne, c'est bien en maintenant la sensation constante que l'agent sait que son capteur est dans une configuration fixe. Bien sûr, si l'environnement voit son état évoluer dans le temps, le raisonnement précédent peut ne plus s'appliquer : ce sera d'ailleurs un des points que nous étudierons plus en détails en §3.3.4. Néanmoins, introduire la notion de transformation compensable de l'environnement permet justement formellement de traiter ce problème, avec le soucis que celles-ci peuvent également détecter des caractéristiques compensables non spatiales. Afin d'expliciter un peu mieux les grandeurs en jeux dans le raisonnement, la sous section suivante introduit les concepts et notation nécessaires à l'introduction des premiers travaux ayant traité cette estimation de la dimension de l'espace.

1. Pour rappel, un groupe est un ensemble doté d'une loi de composition interne pour laquelle il existe un élément neutre, et vérifiant les propriétés d'associativité et de symétrie.

3.2.1.2 Notations et hypothèses

Dans toute la suite, nous allons considérer un agent naïf qui peut interagir avec son environnement par l'application de commandes motrices appartenant à un ensemble des configuration motrices $\mathcal{M}$. Cet ensemble peut être décrit par les variables latentes qui paramétrisent les états de ses actionneurs, représentés par la configuration motrice $\boldsymbol{m} \in \mathcal{M}$. L'agent est également doté de capteurs rigides placés sur différentes parties de son corps qui l'informent sur l'état physique de l'environnement via sa configuration sensorielle $\boldsymbol{s} \in \mathcal{S}$, où $\mathcal{S}$ désigne l'ensemble des configurations sensorielles. Les configurations motrices $\boldsymbol{m}$ et sensorielles $\boldsymbol{s}$ constituent le flux sensorimoteur, seule donnée accessible à l'agent. Bien sûr, la configuration sensorielle de l'agent dépend de sa configuration motrice, de sorte que

$$\boldsymbol{s} = \Psi(\boldsymbol{m}, \boldsymbol{\epsilon}) = \Psi_{\boldsymbol{\epsilon}}(\boldsymbol{m}), \tag{3.1}$$

où $\Psi_{.}(.)$ désigne la *loi sensorimotrice* capturant les propriétés de l'interaction de l'agent avec son environnement, et où $\boldsymbol{\epsilon} \in \mathcal{E}$ représente la configuration de l'environnement, avec $\mathcal{E}$ l'ensemble des configurations de l'environnement.

Dans un premier temps, nous supposerons que les configurations motrices sont définies par des vecteurs $\boldsymbol{m} = (m_1, \ldots, m_{N_m})^T \in \mathcal{M} = \mathbb{R}^{N_m}$, avec N_m le nombre d'actionneurs de l'agent. De la même façon, sa configuration sensorielle $\boldsymbol{s}$ est donnée par ses N_s capteurs extéroceptifs, avec $\boldsymbol{s} = (s_1, \ldots, s_{N_s})^T \in \mathcal{S} \subset \mathbb{R}^{N_s}$. Enfin, nous envisageons que la configuration de l'environnement $\boldsymbol{\epsilon}$ peut être totalement décrit par N_e variables (possiblement non indépendantes), de sorte que $\boldsymbol{e} = (e_1, \ldots, e_{N_e})^T \in \mathcal{E} = \mathbb{R}^{N_e}$. La relation sensorimotrice $\boldsymbol{s} = \Psi_{\boldsymbol{\epsilon}}(\boldsymbol{m})$ nous indique que l'ensemble des sensation $\mathcal{S}$ sont générées depuis les espaces $\mathcal{M}$ et $\mathcal{E}$, mais en aucun cas nous pouvons dire que $\mathcal{S} = \mathbb{R}^{N_s}$. En effet, la nature de l'interaction de l'agent avec son environnement contraint les configurations sensorielles sur un sous-espace $\mathcal{S}$ de $\mathbb{R}^{N_s}$, possiblement courbe et de dimension inférieure à N_s. Notons d le nombre de variables indépendantes permettant de caractériser cette interaction. Nous pouvons alors faire l'hypothèse que $\mathcal{S}$ est une sous-variété différentielle issue d'un plongement de $\mathbb{R}^d$ dans $\mathbb{R}^{N_s}$, et d correspond alors à sa dimension intrinsèque. Nous définirons plus formellement ce que capture précisément cet espace des variables latentes dans la sous section 3.3.4.

3.2.1.3 L'approche de Philipona

C'est au début des années 2000 qu'une première formalisation mathématique des idées de Poincaré est apparue (PHILIPONA, J. K. O'REGAN et NADAL, 2003). Ce travail visait à montrer comment des organismes biologiques sont à même de percevoir des notions de "corps", "environnement", "objet", etc. sur la seule base du flux sensorimoteur. La première problématique abordée concerne un aspect primordial de l'espace : sa dimension. Si le flux sensorimoteur s'appuie vraisemblablement sur des données à très haute dimension, la dimension *a priori* réduite de l'espace géométrique invite à caractériser les invariants de l'interaction avec l'environnement dans un nombre réduit de paramètres. Et en particulier il semble important, avant tout travail visant à définir les notions listées plus haut, de vérifier que la tridimensionalité de l'espace est bien une propriété que l'on peut retrouver au sein des dépendances sensorimotrices (PHILIPONA, 2008). S'appuyant sur l'interaction de l'agent avec son environnement formalisée via la loi sensorimotrice, il est proposé dans (PHILIPONA, J. K. O'REGAN et NADAL, 2003) de définir formellement la notion de compensabilité via la confrontation des deux origines différentes possibles des variations de sensation de l'agent : le mouvement de l'agent et les variations de configuration de l'environnement. Ainsi, en un point particulier de l'expérience sensorimotrice $\boldsymbol{s}_0 = \psi_{\boldsymbol{\epsilon}_0}(\boldsymbol{m}_0)$, nous pouvons exprimer localement les variations sensorielles infinitésimales $\mathrm{d}\boldsymbol{s}|_{(\boldsymbol{m}_0, \boldsymbol{\epsilon}_0)}$ autour de $\boldsymbol{s}_0$ au

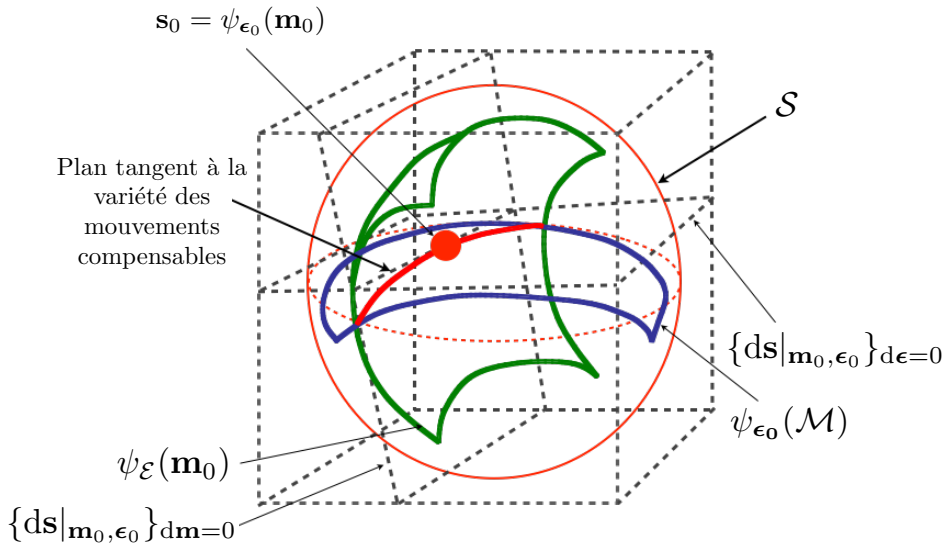


FIGURE 3.3 – Illustration schématique du raisonnement proposé dans (PHILIPONA, J. K. O'REGAN et NADAL, 2003). L'intersection non nulle entre les 2 variétés sensorielles obtenues lorsque seul l'environnement change ou seules les configurations motrices changent représente les variations sensorielles compensables (en rouge gras). L'estimation de sa dimension est effectuée via son plan tangent au point de fonctionnement, de même dimension que la variété. Figure tirée de (GAS et ARGENTIERI, 2016).

premier ordre selon

$$\mathrm{d}\boldsymbol{s}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0} = \frac{\partial \Psi}{\partial \boldsymbol{m}}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0}\mathrm{d}\boldsymbol{m} + \frac{\partial \Psi}{\partial \boldsymbol{\epsilon}}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0}\mathrm{d}\boldsymbol{\epsilon}. \tag{3.2}$$

Nous faisons ainsi apparaître l'espace tangent $\{\mathrm{d}\boldsymbol{s}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0}\}$ à la sous-variété sensorielle $\mathcal{S}$ en une configuration particulière de l'expérience sensorimotrice $\boldsymbol{s}_0, \boldsymbol{m}_0, \boldsymbol{\epsilon}_0$, comme étant engendré par les 2 sous-variétés $\{\mathrm{d}\boldsymbol{s}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0}\}_{\mathrm{d}\boldsymbol{\epsilon}=0}$ et $\{\mathrm{d}\boldsymbol{s}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0}\}_{\mathrm{d}\boldsymbol{m}=0}$, de sorte que

$$\{\mathrm{d}\boldsymbol{s}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0}\} = \{\mathrm{d}\boldsymbol{s}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0}\}_{\mathrm{d}\boldsymbol{\epsilon}=0} + \{\mathrm{d}\boldsymbol{s}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0}\}_{\mathrm{d}\boldsymbol{m}=0}. \tag{3.3}$$

Du fait de l'approximation au premier ordre, $\{\mathrm{d}\boldsymbol{s}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0}\}_{\mathrm{d}\boldsymbol{\epsilon}=0}$ et $\{\mathrm{d}\boldsymbol{s}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0}\}_{\mathrm{d}\boldsymbol{m}=0}$ sont en pratique les deux plans tangents aux deux variétés correspondant respectivement aux variations sensorielles lorsque l'environnement est fixe et lorsque les configurations motrices sont fixes. Toutes ces variétés sont représentées schématiquement à la figure 3.3, qui met en évidence que l'intersection de ces deux variétés n'est pas nulle. En pratique, cette intersection représente les variations sensorielles autour de $\boldsymbol{s}_0$ qui peuvent être produites autant par un changement des configurations motrices de l'agent que par un changement de configuration de l'environnement. Cette sous-variété représente donc l'interaction de l'agent (via son action) avec son environnement et correspond justement aux variations compensables censées la caractériser. Dès lors, déterminer la dimension intrinsèque de cette sous-variété permet d'estimer le nombre de variables latentes qui paramétrisent cette interaction. La propriété de transversalité nous permet d'écrire

$$\begin{aligned}\dim\left(\{\mathrm{d}\boldsymbol{s}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0}\}\right) = \dim\left(\{\mathrm{d}\boldsymbol{s}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0}\}_{\mathrm{d}\boldsymbol{\epsilon}=0}\right) + \dim\left(\{\mathrm{d}\boldsymbol{s}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0}\}_{\mathrm{d}\boldsymbol{m}=0}\right) - \\ \dim\left(\{\mathrm{d}\boldsymbol{s}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0}\}_{\mathrm{d}\boldsymbol{\epsilon}=0} \cap \{\mathrm{d}\boldsymbol{s}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0}\}_{\mathrm{d}\boldsymbol{m}=0}\right)\end{aligned} \tag{3.4}$$

ou encore

$$b = m + e - d. \tag{3.5}$$

On pourra noter que les dimensions de b, m et e ne sont pas forcément égales à $N_m + N_e$, N_m et N_e respectivement : dans le cas général, l'agent et son interaction peuvent présenter des redondances qui vont tendre à diminuer ces valeurs. Nous allons maintenant voir dans la

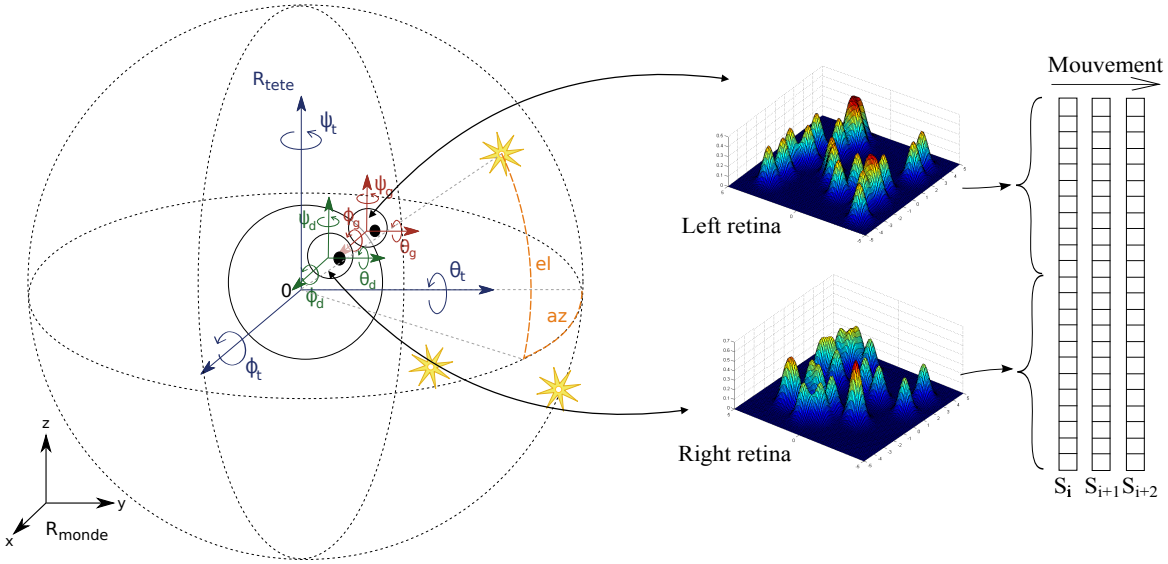


FIGURE 3.4 – Un agent doté d'une tête et deux yeux mobiles en rotation perçoit son environnement constitué de sources lumineuses ponctuelles, placées à une distance constante de l'agent. Pour différentes phases de l'exploration, l'agent capture les vecteurs sensoriels obtenus, qui sont par la suite analysés pour estimer la dimension de la variété sur laquelle ils se trouvent. Figure inspirée de (LAFLAQUIÈRE, 2013).

suite comment estimer ces dimensions et les limites conceptuelles de cette approche, limites que nous allons chercher à dépasser.

3.2.2 Reprise des travaux de Philipona

Un des premiers objectifs de la thèse d'A. LAFLAQUIÈRE a consisté à reprendre les travaux précédents (PHILIPONA, J. K. O'REGAN et NADAL, 2003) et à les appliquer à un robot plus "réaliste" afin de juger de la pertinence et de l'applicabilité du formalisme proposé. Ainsi, là où un bras articulé muni de deux capteurs visuels sont originellement utilisés, nous avons proposé d'exploiter à la place un agent (toujours simulé) doté d'une tête sphérique munie de deux yeux, comme représenté à la figure 3.4. Doté de degrés de redondance motrice supplémentaires, l'agent proposé nous a permis dans un premier temps de mettre en évidence les limites expérimentales du formalisme linéaire initial.

En pratique, la tête comme les deux yeux de l'agent proposé sont capables de bouger indépendamment les uns des autres selon 3 rotations, pilotées par l'intermédiaire de $N_m = 9$ commandes motrices. Les sensations sont issues des deux yeux, au sein desquelles sont placées deux rétines munies de 20 cellules sensibles réparties aléatoirement. Le vecteur de sensation ainsi obtenu est donc de dimension $N_s = 40$. Enfin, l'environnement est constitué de 3 sources lumineuses ponctuelles situées à une distance constante de l'agent, de sorte que chacune de leurs positions est fixée par 2 paramètres indépendants. Ainsi, la configuration de l'environnement est totalement décrite par $N_e = 6$ paramètres. Sur cette base, les données sensorielles qui seront analysées pour en déterminer la dimension intrinsèque sont générées de la façon suivante :

— une configuration sensorimotrice "de référence" $\boldsymbol{s}_0, \boldsymbol{m}_0, \boldsymbol{\epsilon}_0$ est tirée au sort[2] ;

2. La génération aléatoire de ces données est tout de même guidée pour faire en sorte que l'agent puisse voir les sources lumineuses.

— trois matrices sensorielles S, issues de 3 explorations différentes, sont simulées : (i) seul l'agent bouge, (ii) seulement l'environnement bouge, et (iii) l'agent et l'environnement bougent tous les deux. A cette occasion, $N = 1000$ mouvements aléatoires d'amplitude A sont simulés, produisant ainsi une matrice S de dimension $N_s \times N$;

— l'expérience est répétée 100 fois de façon à analyser statistiquement la simulation ;

— enfin, l'ensemble du processus est répété pour différentes amplitudes A des mouvements de rotation simulés.

A l'issue de la simulation, et pour une amplitude et une génération aléatoire des mouvements, nous disposons donc de 3 matrices sensorielles différentes dont les vecteurs vivent sur les trois variétés $\{\mathrm{d}\boldsymbol{s}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0}\}$, $\{\mathrm{d}\boldsymbol{s}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0}\}_{\mathrm{d}\boldsymbol{\epsilon}=0}$ et $\{\mathrm{d}\boldsymbol{s}|_{\boldsymbol{m}_0,\boldsymbol{\epsilon}_0}\}_{\mathrm{d}\boldsymbol{m}=0}$. Leurs dimensions intrinsèques respectives b, m et e peuvent être estimées à l'aide d'une décomposition en valeurs singulières (SVD) des 3 matrices de données : la limite entre leurs valeurs singulières significatives et non significatives permet d'estimer le nombre de dimension minimal expliquant les données sensorielles. Enfin, sur la base des 100 répétitions de l'exploration, nous pouvons déterminer un taux de bonne estimation de la dimension intrinsèque des données sensorielles, en comparant le résultat de la SVD avec les valeurs théoriques des dimensions :

— lorsque seul l'agent bouge, et comme il y a autant de commandes motrices que d'actionneurs, nous avons $m = 9$;

— lorsque seul l'environnement bouge, sa configuration est totalement déterminée par N_e paramètres indépendants, de sorte que nous avons $e = 6$;

— enfin, lorsque l'agent et l'environnement bougent tous les deux, la dimension b ne vaut pas 15 : sa dimension est plus faible du fait de l'existence de variations sensorielles pouvant être générées à la fois par l'environnement ou le mouvement de l'agent, i.e. par des variations compensables. En tant qu'observateur externe, nous savons que les 3 rotations centrées sur la tête correspondent aux seuls mouvements compensables. De fait, nous avons alors théoriquement $d = 3$, et ainsi $b = 12$.

Obtenir $d = 3$ semble indiquer que l'interaction de l'agent avec son environnement s'effectue dans un monde à 3 dimensions. En réalité, d capture le nombre de variables indépendantes qui caractérisent cette interaction : si jamais l'agent et les sources lumineuses de l'environnement pouvaient se translater dans les 3 directions de l'espace, alors nous aurions obtenus $d = 6$: 3 translations et 3 rotations paramétrisent l'interaction. En pratique, c'est bien la dimension du groupes des transformations compensables qui est estimé, et plus précisément celle du groupe de Lie des transformations orthogonales (POINCARÉ, 1895).

Les taux de bonne estimation de ces dimensions sont reportés sur la figure 3.5, pour différentes amplitudes des mouvements de rotation. On peut y constater que l'estimation de d est correcte pour des mouvements angulaires d'amplitudes allant jusqu'à environ 10^{-4} degrés. Au delà, les dimensions des 3 variétés ne sont plus correctement estimées. Il est clair que l'équation (3.3) n'est valable qu'à l'ordre 1, i.e. sur les plans tangents des variétés d'origine. Or, dès que l'amplitude des mouvements augmentent, les données sensorielles ne peuvent plus être correctement représentées par ce plan tangent, dont l'estimation de dimension repose sur une SVD, i.e. une méthode linéaire. Dès lors, les performances d'estimation des dimensions des variétés se réduisent. En pratique, les variétés en jeux sont certainement courbes, de sorte que l'application d'une SVD ne permet pas d'en estimer la dimension intrinsèque. Il est donc clair que l'applicabilité de ces travaux à la robotique est limitée car il est totalement inenvisageable de travailler sur des amplitudes de mouvements aussi faibles. A ce stade, la question est même posée de savoir si tout le raisonnement précédent reste valide au sein d'agents dotés de capacités de mouvement réalistes.

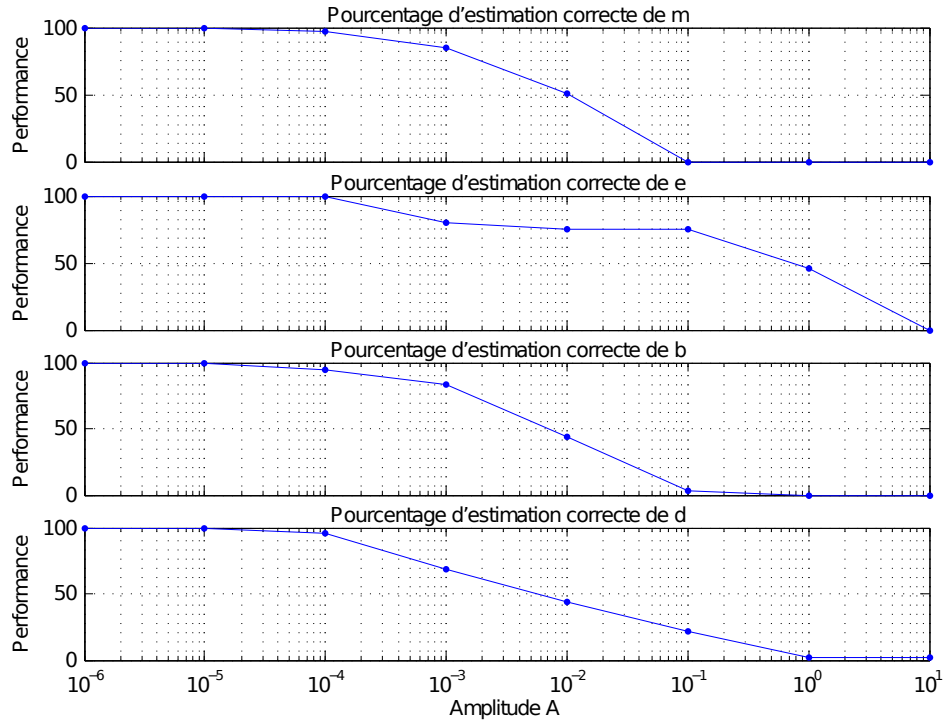


FIGURE 3.5 – Taux de bonne estimation des dimensions m, e et b en fonction de l'amplitude angulaire des mouvements. De ces trois valeurs estimées, on peut déduire $\hat{d} = \hat{m} + \hat{e} - \hat{b}$ et son taux de bonne estimation. Figure tirée de (LAFLAQUIÈRE, ARGENTIERI, BREYSSE et al., 2012).

3.2.3 Extension aux mouvements réalistes

Suite à cette évaluation de l'influence de l'amplitude des mouvements sur l'estimation des dimensions, nous avons orientés les travaux d'A. LAFLAQUIÈRE sur la prise en compte du caractère probablement non-linéaire des variétés sous-jacentes. Il s'agissait alors d'étudier si des méthodes d'analyse non linéaire permettent d'étendre la portée des résultats précédents.

Il apparaît donc que l'approximation linéaire précédente n'est valide qu'au voisinage immédiat du point de fonctionnement sensorimoteur. Pourtant, l'équation (3.3) peut se réécrire d'une manière plus générale en faisant apparaître les 2 variétés "complètes" obtenues lorsque seules les configuration motrices de l'agent changent ($\mathcal{S}_M = \psi_{\boldsymbol{\epsilon}_0}(\mathcal{M})$) et lorsque seul l'environnement change ($\mathcal{S}_E = \psi_{\mathcal{E}}(\boldsymbol{m}_0)$), cf. figure 3.3 , sous la forme

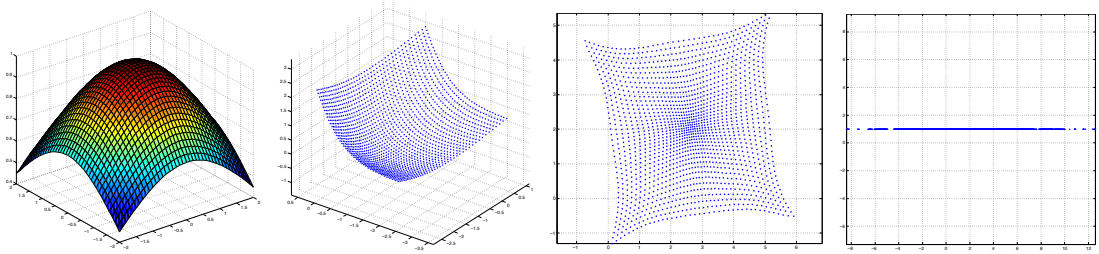
$$\mathcal{S} = \mathcal{S}_M + \mathcal{S}_E. \tag{3.6}$$

La même propriété de transversalité donne alors

$$\dim(\mathcal{S}) = \dim(\mathcal{S}_M) + \dim(\mathcal{S}_E) - \dim(\mathcal{S}_M \cap \mathcal{S}_E) \tag{3.7}$$

$$\Leftrightarrow b = m + e - d. \tag{3.8}$$

Ainsi, les équations obtenues pour l'estimation de dimension restent identiques, mais s'appliquent non plus aux plans tangents des variétés, mais plutôt aux variétés elles-même. Il s'agit alors de déterminer une méthode adéquate d'estimation de leurs dimensions intrinsèques pour espérer être capable à nouveau de déterminer correctement d pour des mouvements de plus grande amplitude.



(A) Données initiales (B) Projection 3D : J(3)=0 (C) Projection 2D : J(2)=84 (D) Projection 1D : J(1)=8416

FIGURE 3.6 – Illustration de l'augmentation de l'erreur de projection en fonction de la dimension cible. Si celle-ci est inférieure à la dimension intrinsèque des données d'entrée, l'erreur J augmente significativement. Figure tirée de (LAFLAQUIÈRE, ARGENTIERI, BREYSSE et al., 2012).

3.2.3.1 Sur l'estimation de la dimension intrinsèque d'une variété

La littérature propose plusieurs solutions à ce problème (LAFLAQUIÈRE, 2013), parmi lesquelles les approches fractales (THEILER, 1990), les méthodes spectrales (LEE et VERLEYSEN, 2007), ou les méthodes par essai-erreurs. Cette dernière approche est plus coûteuse que les précédentes d'un point de vue computationnel ; c'est néanmoins l'approche que nous sélectionnons, car elle ne nécessite pas d'a priori particulier sur la distribution des données à analyser. Ce type d'approche s'appuie sur une technique (a priori quelconque) de projection des données haute-dimension dans un espace de dimension inférieure. Tant que la dimension de projection p est suffisante pour expliquer les données (et donc les projeter en respectant leur topologie, i.e. leurs relations de voisinage), alors c'est que cette projection s'opère à une dimension supérieure ou égale à la dimension intrinsèque des données. L'erreur de projection $J(p)$ est donc faible. Par contre, dès que cette projection est effectuée à une dimension p inférieure strictement à la dimension intrinsèque, le nuage projeté est associé à une erreur de projection $J(p)$importante. Cette augmentation importante de l'erreur de projection permet alors d'estimer la dimension intrinsèque $\widehat{\dim}$ selon

$$\widehat{\dim} = \arg\max_{p} \frac{J(p-1)}{J(p)}. \tag{3.9}$$

Cette idée est illustrée à la figure 3.6 pour un nuage de point 2D immergé dans un espace 3D, et projeté successivement en 3D, 2D et 1D : l'erreur de projection J explose pour $p = 1$ [3]. Reste maintenant à choisir une méthode de projection des données. Dans toute la suite, nous avons opté pour une analyse en composantes curvilignes (CCA) (DEMARTINES et HERAULT, 1997), méthode de réduction de dimension non linéaire basée sur la conservation des faibles distances euclidiennes au moment de la projection. Cette propriété fait de la CCA une méthode particulièrement bien adaptée au traitement de variétés fortement courbées le tout en présentant une complexité algorithmique raisonnable. En pratique, elle consiste à minimiser l'erreur de projection J, définie par

$$J = \frac{1}{2} \sum_{i,j=1}^{N} f(\lambda, d(\boldsymbol{y}_i, \boldsymbol{y}_j)) \left(d(\boldsymbol{s}_i, \boldsymbol{s}_j) - d(\boldsymbol{y}_i, \boldsymbol{y}_j)\right)^2, \tag{3.10}$$

3. On remarquera néanmoins que l'application de (3.9) ne fonctionnerait pas ici. La très faible dimension des données conduit, sur les 2 valeurs possibles à calculer du ratio d'erreur, à une valeur infinie. Ce cas ne se présente pas en pratique pour les données à traiter.

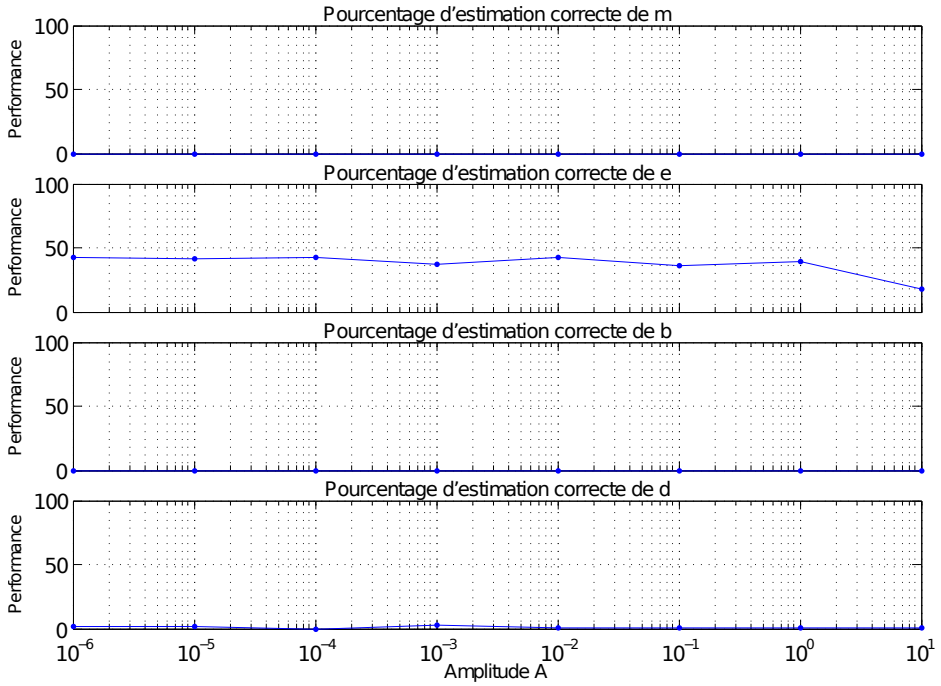
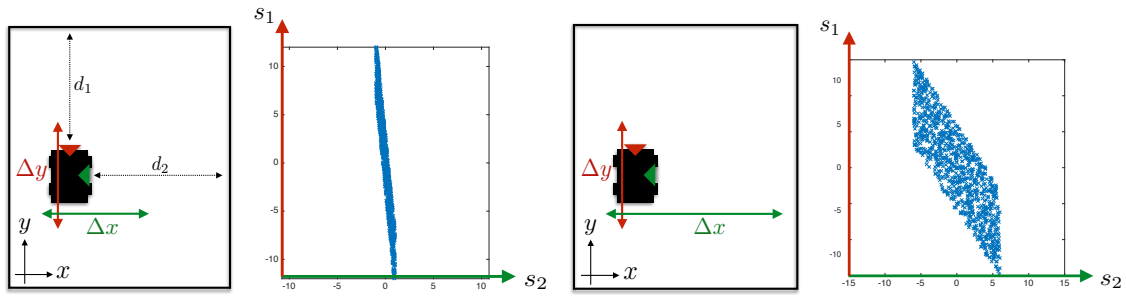


FIGURE 3.7 – Taux de bonne estimation des dimensions m, e et b en fonction de l'amplitude angulaire des mouvemente, obtenu via une CCA sur les données sensorielles. Figure tirée de (LAFLAQUIÈRE, 2013).

où $d(.,.)$ représente la distance Euclidienne entre 2 vecteurs, $\boldsymbol{s}_i$ représente les données d'entrée (à haute dimension), $\boldsymbol{y}_i$ les même données exprimée dans une dimension inférieure p, et $f(\lambda, d(.,.))$ une fonction de voisinage permettant de privilégier la conservation des faibles distances selon un paramètre de voisinage λ. La minimisation de l'erreur J est réalisée par un algorithme de type "descente de gradient", légèrement optimisé pour éviter de nombreux minima locaux. Parmi les méthodes de projection des données, on peut également nommer ISOMAP (TENENBAUM, SILVA et LANGFORD, 2000), travaillant cette fois sur la conservation des distances de graphe, mais sensiblement plus coûteuse en terme de temps de calcul.

3.2.3.2 Application à la variété sensorielle

Tout semble maintenant prêt pour l'application de la CCA à la variété sensorielle. Les données utilisées sont exactement les même que celles exploitées par l'approche linéaire et générées à l'aide de l'agent simulé représenté figure 3.4. La dimension maximale testée pour la projection est $p_{\max} = 15$. Les résultats d'estimation des dimensions m, e, b et donc d sont représentés sur la figure 3.7. De manière surprenante, les résultats obtenus à l'aide de la méthode non linéaire CCA de projection des données n'améliore pas du tout les résultats, même (et surtout) pour des mouvements d'amplitude importante. Appliquer à la place de la CCA une méthode ISOMAP ne change pas la nature des résultats, qui restent en particulier aussi mauvais pour les faibles amplitudes de mouvement : clairement, la courbure des variétés n'est pas la cause de ces mauvaises performances. L'étude attentive des propriétés de la variété sensorielle pointe vers un soucis lié à l'asymétrie des données, c'est à dire un étirement du nuage de données dans une ou plusieurs directions, au détriment d'autres néanmoins aussi porteuses d'information. Ce phénomène est illustré aux figures 3.8a et 3.8b, pour un robot capable de se mouvoir dans les 2 direction x et y et équipé de deux capteurs d'obstacle le renseignant sur les deux distances d_1 et d_2 aux murs environnant. Si l'exploration motrice est effectuée de manière homogène selon les deux directions Δx et Δy, la



(A) Robot dans son environnement (B) Variété sensorielle (C) Ré-échantillonnage moteur (D) Variété sensorielle

FIGURE 3.8 – Illustration de l'étirement des données sensorielles (B), pour un robot mobile dont un des capteurs amplifie sensoriellement une des 2 directions de commande (A). Après ré- échantillonnage moteur visant à amplifier les mouvements selon la direction x (C), la variété sensorielle voit son asymétrie réduite (D). Figure tirée de (GAS et ARGENTIERI, 2016).

variété sensorielle –reliée à la variété motrice par la loi sensorimotrice– présente un étirement important : un des deux capteurs fournit des données variant beaucoup plus que l'autre. L'application d'une CCA (ou même d'une méthode linaire) sur ces données sensorielles donnerait vraisemblablement une dimension du nuage de point égale à 1, alors que sa dimension intrinsèque est de 2. L'analyse d'une telle variété nécessite donc de modifier la façon dont les données sensorielles sont réparties. Nous proposons pour cela une approche de ré-échantillonnage actif des données sensorielles permettant de réduire l'asymétrie du nuage de points.

3.2.3.3 Mise en œuvre d'un ré-échantillonnage moteur

Les méthodes non linéaires d'estimation de dimension sont très sensibles à la répartition des données dans leur espace d'origine. Si celle-ci est très étirée, alors la plupart des méthodes échoueront à déterminer la dimension intrinsèque de la variété sous-jacente. Une première idée était de travailler les données préalablement à leurs projections, typiquement via un centrage et une normalisation. Néanmoins, ces 2 opérations n'ont pas d'effet sur l'asymétrie. Une autre idée pouvait consister à chercher les directions principales du nuage de points, portant le plus de variance. Typiquement, si on oublie la nature possiblement non linéaire du nuage de points, une PCA permettrait de détecter ces dimensions faiblement représentées, associées à des valeurs propres faibles. Seulement, comme nous pouvons nous en douter, il est difficile de différencier une faible variance liées aux données d'une faible variance liée à la présence de bruit. Dès lors, amplifier ces composantes pourrait revenir à amplifier le bruit. Il est donc absolument nécessaire d'impliquer le comportement exploratoire de l'agent dans la regénération des données sensorielles. En d'autres termes, c'est l'agent qui doit, itérativement et de proche en proche, déterminer comment explorer son environnement de façon à doter sa variété sensorielle des bonnes propriétés permettant l'estimation correcte de sa dimension intrinsèque. Pour cela, nous avons proposé d'appliquer la stratégie de rééchantillonnage moteur suivante , dont les principales étapes successives sont :

1. *Détermination des faibles variances sensorielles*, via une décomposition en valeurs singulières de la matrice des données sensorielles ;
2. *Détermination des commandes motrices associées aux variances sensorielles*, par un changement de base des N_s premiers vecteurs singuliers à droite ;
3. *Suppression des redondances dans la loi sensorimotrice*, là aussi en exploitant les $N - N_s$ vecteurs singuliers à droite ne générant aucune variation sensorielle ;

4. *Exploitation des valeurs numériques des valeurs singulières* pour amplifier les directions faiblement représentées au sein des données grâce à la base obtenue précédemment ;
5. Enfin, on exprime les commandes motrices dans leur base d'origine, en veillant à les remettre à l'amplitude A de mouvement demandée.

Ces différents points sont répétés itérativement, après une exploration initiale aléatoire et de même variance selon toutes les dimensions. Une fois arrivé à l'étape 5, l'agent réalise les mouvements ainsi déterminés, produisant alors un nouveau nuage de points analysé ensuite en repartant de l'étape 1. Appliqué à l'exemple précédent du robot mobile, nous pouvons imaginer que le nuage de points sensoriels, étiré, représenté sur la figure 3.8b puisse être déformé de façon à réduire son asymétrie en augmentant les mouvement du robot selon la commande motrice Δx, cf. figure 3.8c et 3.8d. Et il est maintenant beaucoup plus probable qu'une méthode d'estimation de la dimension intrinsèque de la variété sensorielle puisse fournir de bons résultats. Évidemment, dans un cas plus complexe pour lequel les dimensions des variétés en jeux sont bien supérieures, il est particulièrement difficile de représenter visuellement le bénéfice de ce ré-échantillonnage. Mais il est attendu que le taux d'estimation de la dimension augmente significativement.

Avant d'évaluer ce gain, il est important de noter que la stratégie de ré-échantillonnage telle que proposée s'appuie sur une décomposition en valeurs singulières de la matrice de sensations. Elle ne devrait donc pouvoir s'appliquer qu'à des variétés faiblement courbées, ce qui n'est a priori pas le cas des données sensorielles obtenues avec de grandes amplitudes de mouvement. Nous proposons deux manières de l'appliquer néanmoins :

— *locale* : le ré-échantillonnage est appliqué uniquement aux données issues d'un mouvement d'amplitude infinitésimale. Le résultat de ce ré-échantillonnage est ensuite amplifié à l'amplitude A demandée : cette dernière amplification des commandes motrices obtenues localement à des mouvement plus grands ne garantit en rien que la loi sensorimotrice ne va pas à nouveau étirer significativement les données sensorielles ;

— *globale* : le ré-échantillonnage est appliqué directement aux données issues d'un mouvement d'amplitude A quelconque : selon la courbure des données, il n'y a donc aucune garantie de convergence vers un résultat satisfaisant.

L'estimation des dimensions intrinsèques des différentes variétés est à nouveau représentée sur la figure 3.9 pour ces 2 stratégies. On peut y voir que la stratégie de ré-échantillonnage locale permet d'envisager des mouvements de quelques degrés, là où l'approche linéaire seule ne permettait que des mouvements infinitésimaux. Néanmoins, si on envisage son utilisation globale, alors les performances chutent d'un facteur 10, et seuls des mouvements d'environ $0.1°$ sont envisageables. C'est toujours bien mieux qu'avec l'approche linéaire (gain d'un facteur 1000), et cet ordre de grandeur est maintenant physiquement atteignable (mais reste difficile à mettre en œuvre). Clairement, l'utilisation d'une méthode linéaire au sein du rééchantillonnage moteur en dégrade les performances, et l'utilisation d'une méthode intrinsèquement non-linéaire devrait permettre d'en améliorer la portée.

Publication

Le travail synthétisé dans cette sous-section a donné lieu à l'article (LAFLAQUIÈRE, ARGENTIERI, BREYSSE et al., 2012) publié et présenté au sein de la conférence IEEE IROS.

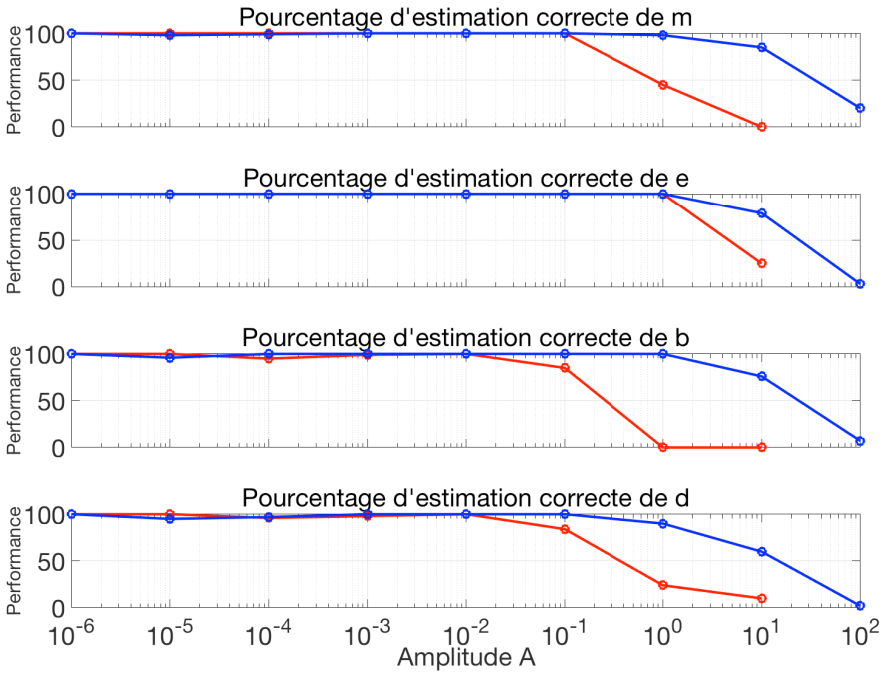


FIGURE 3.9 – (bleu) .

3.2.4 Discussion et conclusion

Les travaux précédents avaient pour objectifs de montrer qu'il était possible d'extraire, au sein du flux sensorimoteur, une information a priori triviale : la dimension de l'interaction d'un agent naïf avec son environnement. L'extraction de cette dimension, appelée par abus de langage "dimension de l'espace", permet de mettre en évidence que malgré les dimensions importantes du flux sensorimoteur, l'interaction pouvait être caractérisée par un nombre réduit de paramètres, capturant ici des informations spatiales. Alors que jusqu'à présent cette information n'était accessible qu'à une échelle infinitésimale, la stratégie de rééchantillonnage moteur proposée permet d'atteindre une échelle atteignable pour des agents robotiques : l'intuition de Poincaré semble donc bien accessible à des agents aux capacités motrices proches de celles dont sont dotées nos robots. Ce rééchantillonnage est par ailleurs une approche intrinsèquement active, dans la mesure où les données sensorielles sont systématiquement régénérées à chacune des étapes de l'exploration motrice de l'agent : la loi sensorimotrice est donc sollicitée à chaque itération. D'une certaine façon, nous voyons donc apparaître ici une certaine finalité de l'action, dont l'objectif ici pourrait être de caractériser l'interaction de l'agent dans le but de construire un sens de la perception.

Néanmoins, certaines limites conceptuelles viennent nuancer la portée des résultats obtenus. Tout d'abord, la façon dont l'exploration motrice est conduite reste naïve. Pour rappel, celle-ci s'appuie sur trois collectes sensorielles dans le cas où (i) seul le robot bouge, (ii) seul l'environnement bouge, et (iii) les deux bougent. Le cas (i) est problématique : comment l'agent, de son point de vue, peut-il s'assurer qu'il est le seul à bouger alors qu'il n'a aucun contrôle sur l'environnement ? Plusieurs solutions peuvent être envisagées : il peut être argumenté que l'exploration motrice de l'agent s'effectue à une dynamique plus rapide que celle de l'environnement. Une autre solution pourrait consister pour l'agent à revenir régulièrement en sa configuration motrice de référence $\boldsymbol{m}_0$ pour s'assurer que la sensation en cette configuration n'a pas changé. Son exploration serait alors de type saccadique : on retrouve quelque part cette hypothèse sur la dynamique de l'exploration, alors que cet aspect dynamique n'a jamais été pris en compte. Il s'agirait certainement de modifier la loi sensorimotrice en une équation différentielle pour rendre compte de ce phénomène. Par

ailleurs, nous savons que la dimension extraite de ces explorations correspond en pratique au groupe des transformations compensables. Or, ces compensations n'ont été envisagées que de manière totale : la sensation complète doit pouvoir être retrouvée à l'identique par une action de l'agent après un changement de configuration de l'environnement. Cela n'est pour l'instant envisageable que pour des environnements simples, impliquant par exemple seulement quelques sources de lumières ponctuelles. Dans le cas d'un environnement réaliste, fait d'objets rigides dotés de textures complexes, compenser en totalité la sensation n'est tout simplement pas réalisable.

Au final, il est important de comprendre que l'extraction de la dimension de l'interaction de l'agent avec son environnement n'est pas une finalité en soit. Nous n'avons certainement pas au sein de notre cerveau un groupe de neurones dédiés à cette tâche, clamant en permanence "$d = 3$". Il s'agissait avant tout de montrer en quoi le flux sensorimoteur contient, en son sein, une telle information. Nous allons maintenant voir dans la section suivante que d'autres types de structures peuvent en être extraites.

3.3 Extraire une structure des invariants sensorimoteurs

La section précédente était dédiée à l'estimation du nombre minimal de paramètres caractérisant l'interaction d'un agent naïf avec son environnement. Il est clair que cette information, aussi critique soit-elle, ne permet pas encore de montrer que l'agent a su capturer, au sein de son flux sensorimoteur, des structures invariantes pouvant être à l'origine de sa compréhension du monde. Nous avons souhaité que la seconde partie de la thèse d'Alban LAFLAQUIÈRE s'attaque en partie à ce problème, avec pour objectif de proposer l'élaboration graduelle d'une formalisation sensorimotrice de la perception. A. LAFLAQUIÈRE a eu l'intuition que la structure du flux sensorimoteur était susceptible de prendre racine dans ce que nous avons appelé naïvement des courbes noyaux, déterminées dans l'espace des configurations motrices de l'agent. Cette intuition a été confirmée par des résultats expérimentaux obtenus en simulation ; ils sont synthétisés dans la sous-section suivante. Pour autant, aucune formalisation mathématique solide de l'approche n'avait été proposée à ce stade. Nous avons ainsi souhaité travailler sur la formalisation précise de cette approche à l'occasion de la thèse de Valentin MARCEL. En introduisant les concepts de la thèse d'A. LAFLAQUIÈRE sous une forme plus générale d'ensembles quotient, V. MARCEL a démontré le lien formel existant entre ces ensembles et ce qu'ils représentent : un espace abstrait caractérisant l'interaction de l'agent. La formalisation précise proposée, ainsi que son exploitation pour la découverte du corps d'un agent est proposée en §3.3.3. Enfin, si la restriction du problème au corps de l'agent permet de simplifier son étude et sa formalisation, nous avons souhaité travailler à étendre ces résultats afin de tenir compte de l'évolution de l'état de l'environnement au cours du temps. C'est précisément l'objectif de la fin de thèse de V. MARCEL que d'intégrer le formalisme sensorimoteur au sein de la vie d'un agent. Cela pourrait permettre d'envisager l'intégration sensorimotrice comme une expérience continue et itérative qui enrichit la représentation de l'interaction au fur et à mesure que l'agent fait l'expérience de nouveaux environnements. Des premiers résultats de ce travail encore en cours sont présentés en §3.3.4.

3.3.1 Approche intuitive

Avant de détailler l'approche que nous envisageons, essayons dans un premier temps de définir quelles propriétés nous cherchons à capturer au sein de l'expérience sensorimotrice. Dans un premier temps, il semble particulièrement pertinent d'espérer doter un agent de capacités d'évolution dans l'espace, de planification d'un mouvement, le tout en s'appuyant uniquement sur le flux sensorimoteur. Cela semblerait en effet témoigner de sa capacité de compréhension de son interaction via un comportement spatial adapté. Comme nous le verrons dans la suite, nous proposons pour cela que l'agent cherche à se doter d'une représentation interne de l'espace dans laquelle il serait capable de définir ces notions, relativement à ses capacités motrices et sensorielles. Cette représentation interne serait donc subjective, comme le suggère l'approche SMC. Cette même approche suggère néanmoins qu'une telle représentation –même relative aux capacités de l'agent– n'est pas nécessaire, voir même inexistante. A vrai dire, plutôt qu'une représentation interne *de l'espace*, nous proposons de travailler sur une représentation interne *des relations sensorimotrices* qui sont susceptibles de décrire l'espace. Et comme nous aurons l'occasion de l'indiquer plus loin, la construction d'une telle représentation nous permettra d'illustrer qu'il est possible de capturer ces relations. Nous n'irons donc pas jusqu'à dire qu'il est obligatoire d'établir une telle représentation afin qu'un agent puisse capturer les caractéristiques de son expérience sensorimotrice.

Quelles doivent être alors les propriétés devant être capturées par cette représentation interne ? Nous faisons ici l'hypothèse qu'il est nécessaire (mais peut être pas suffisant) de conserver la topologie de l'espace externe au sein de cette représentation. Nous allons illustrer cette intuition dans la suite, avant d'en proposer une vision plus formelle.

3.3.1.1 De la variabilité de l'expérience sensorielle

Exactement comme nous l'avions fait dans la section précédente, l'existence a priori de l'espace comme contenant commun à l'agent et son environnement permet d'envisager l'utilisation de la variété sensorielle explorée via les capacités d'action de l'agent comme représentation basse dimension de l'espace. Prenons pour cela l'exemple simple illustré sur la figure 3.10 dans lequel un agent doté d'une seule configuration motrice $\boldsymbol{m} = m$ se déplace linéairement pour atteindre une position x dans l'espace. Cet agent dispose de deux capteurs extéroceptifs générant deux scalaires s_1 et s_2 images des distances moyennes aux sources lumineuses ponctuelles définissant l'environnement. Bien sûr, l'agent n'a accès qu'à son flux sensorimoteur, constitué des données m, s_1 et s_2 : dès lors, comment peut-il comprendre que son interaction spatiale avec l'environnement se résume à ce paramètre x, identifié d'un point de vue externe ? Si on considère uniquement la variété sensorielle, telle que représentée à la figure 3.10, alors il est évident que celle-ci est de dimension intrinsèque égale à 1. De plus, à chacune des valeurs sensorielles correspond une seule et unique valeur de x [4] : la variété sensorielle a bien capturé le seul paramètre caractérisant l'interaction de l'agent avec son environnement. D'ailleurs, de son point de vue, la dimension de l'espace (ou plutôt de son interaction) est de 1, alors que nous avons représenté schématiquement ce même agent dans le plan. Cependant, la variété sensorielle reste dépendante du contenu de l'environnement, et plus généralement de sa configuration $\boldsymbol{\epsilon}$. Pour des sources ponctuelles placées autrement dans l'environnement, la variété sensorielle est certes toujours de dimension 1, mais différente. Cette dépendance à la configuration de l'environnement est problématique et ne semble pas en accord avec la notion d'espace indépendante des objets qu'il contient. On notera que les approches traditionnelles en perception robotique s'appuient malgré tout

4. Il faut néanmoins que l'environnement soit suffisamment riche pour que des cas dégénérés ne se présentent pas, i.e. la variété sensorielle pourrait présenter des réductions locales de dimension dans le cas général.

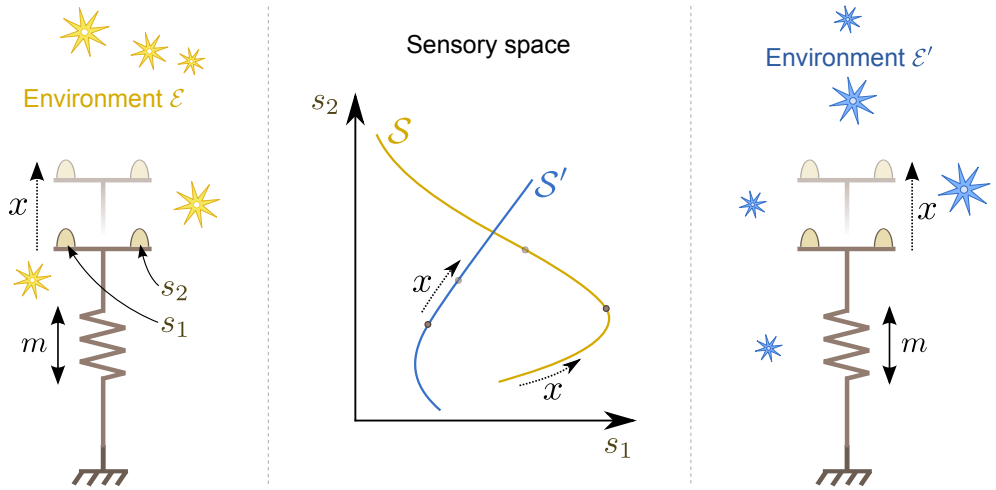


FIGURE 3.10 – Illustration de la dépendance à l'environnement de l'expérience sensorielle. Un agent simple, monodimensionnel, capte au sein de sa variété sensorielle l'unique degré de liberté paramétrant son interaction avec l'environnement. Seulement, si la configuration de l'environnement change, la variété sensorielle est modifiée également. Figure tirée de (LAFLAQUIÈRE, J. K. O'REGAN et al., 2015).

uniquement sur la variété sensorielle, comme nous l'avons nous même faits dans les travaux précédents portant sur l'audition en robotique. Pourtant, la représentation de l'interaction que nous devrions obtenir devrait être identique pour tout état de l'environnement. Travailler uniquement avec la variété sensorielle pose problème ici : il faut donc vraisemblablement inclure la composante motrice du flux sensorimoteur dans le raisonnement.

3.3.1.2 Les ensembles noyaux comme invariants sensorimoteurs

Exemple simple : Reprenons le même exemple que précédemment, mais avec un agent doté cette fois de 2 commandes motrices m_1 et m_2 redondantes, cf. figure 3.11. il est clair que si l'état de l'environnement est le même que celui envisagé au sein de la figure 3.10, la variété sensorielle obtenue avec ce second agent est identique au cas précédent. Cependant, la structure de son espace moteur est différente : de par sa redondance, il existe un sous-ensemble de configurations motrices $\mathcal{M}^{\boldsymbol{s}} \subset \mathcal{M}$ qui génèrent la même sensation $\boldsymbol{s}$, avec

$$\mathcal{M}^{\boldsymbol{s}} = \{\boldsymbol{m} \in \mathcal{M} | \boldsymbol{s} = \Psi_{\boldsymbol{\epsilon}}(\boldsymbol{m})\}. \tag{3.11}$$

Ces ensembles $\mathcal{M}^{\boldsymbol{s}}$ peuvent être vus comme le noyau des fonctions $\Psi_{\boldsymbol{\epsilon}} - \boldsymbol{s}, \forall \boldsymbol{s} \in \mathcal{S}$, et seront donc appelés par abus de langage *ensembles noyaux* dans la suite. Il est important de comprendre que ces ensemble noyaux capturent l'existence de contraintes qui s'appliquent au flux sensorimoteur de l'agent. Ces contraintes sont directement liées à la nature de l'interaction entre l'agent et l'environnement. En effet, si nous faisons l'hypothèse que toutes les sensations $\boldsymbol{s}$ captées par l'agent sont différentes pour une configuration de l'environnement donnée[5], alors ces sensations sont en fait acquises par l'agent en des positions x différentes du capteur extéroceptif dont il est doté. Cela veut donc dire qu'à une sensation $\boldsymbol{s}$ donnée correspond une seule est unique position x dans l'espace, exactement comment dans le cas précédent. Seulement, cette coïncidence peut être maintenant détectée par l'agent par les ensemble $\mathcal{M}^{\boldsymbol{s}}$: si jamais la configuration de l'environnement change, la valeur sensorielle associée à une position x va être modifiée (et c'était justement la limite pointée dans la sous

5. Si ce n'est pas le cas, alors l'interprétation de cette interaction par l'agent pour être différente de la notre, obtenue d'une point de vue externe. Elle reste néanmoins valide du point de vue de l'agent.

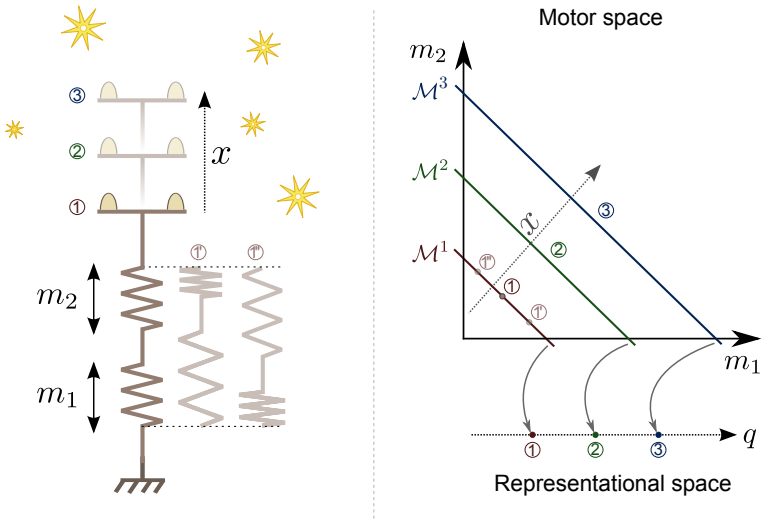
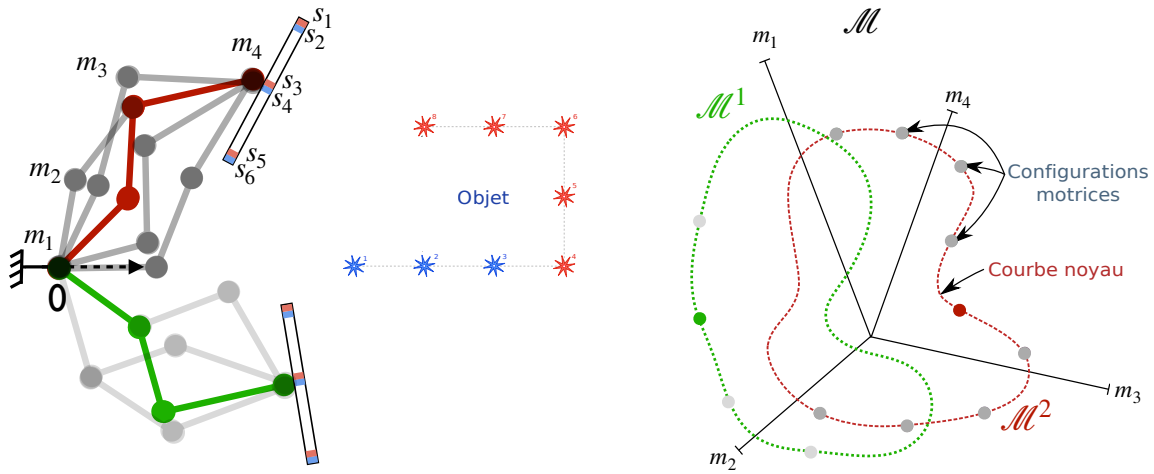


FIGURE 3.11 – L'agent précédent est maintenant doté de deux degrés de liberté pour contrôler son unique mouvement selon x. Cette redondance motrice peut être capturée au sein de son espace moteur, dont la structure de invariants capture le paramètre x au sein d'une représentation interne. Figure tirée de (LAFLAQUIÈRE, J. K. O'REGAN et al., 2015).

section précédent), mais pas l'ensemble $\mathcal{M}^s$: il serait donc plus juste de noter ces ensemble $\mathcal{M}^q$, où q représente les variables latentes caractérisant l'interaction.

Dans le cas illustratif de la figure 3.11, les ensembles noyaux prennent la forme, au sein de $\mathcal{M} = \{(m_1, m_2) \in \mathbb{R}^2\}$, de droites d'équation $x = m_1 + m_2$. Si on dispose d'une mesure de distance entre ces ensembles noyau, alors il est possible de les projeter en de multiples points au sein d'un espace abstrait en respectant leurs distances relatives. Cette opération est représentée sur la figure 3.11 (droite) : chacune des droites est représentée par un paramètre q_i au sein d'une variété monodimensionnelle, et ce paramètre q_i représente de manière équivalente une position x_i du capteur dans l'espace. Cette variété, obtenue uniquement sur la base d'invariants sensorimoteurs, peut donc être utilisée comme représentation interne de l'interaction de l'agent avec son environnement. Cependant, une hypothèse fondamentale doit ici être posée : *à deux ensembles noyau proches doivent correspondre deux configurations sensorielles proches, qui doivent à leur tour correspondre à deux configurations spatiales proches des capteurs dans l'espace*. A ce stade, rien ne permet de comprendre en quoi cette hypothèse est valide, ou du moins ce qu'elle implique éventuellement sur la représentation. Nous serons amenés plus tard à discuter de cette hypothèse, qui grâce à la formalisation présentée plus loin prendra une toute autre importance.

Simulation d'un agent plus complexe : Afin de vérifier si l'intuition consistant à travailler sur la structure des ensembles noyau permet effectivement de capturer l'interaction d'un agent avec son environnement, nous exploitons un système robotique simulé à la structure motrice plus complexe. Celui-ci est représenté sur la figure 3.12a. Il s'agit d'un bras robotique plan constitué de 4 degrés de libertés en rotation, dont la configuration motrice $\boldsymbol{m}$ est fixée par 4 commandes $m_1, \ldots, m_4$ représentant les angles de chacune des articulations. L'espace des configurations motrice $\mathcal{M}$ est donc de dimension 4. Une rétine est fixée sur la dernière articulation, précédée d'une lentille de type sténopé. Dessus sont placés 6 cônes sensibles soit à la couleur bleue, soit à la couleur rouge. La sensation s_i générée par chacun des cônes est la somme des contributions de chacune des sources de lumières colorées dans l'environnement, selon la distance qui sépare le cône de leurs projections sur la rétine. La dimension de l'espace des sensations $\mathcal{S}$ est donc de 6. L'environnement est constitué d'un objet rigide, constitué d'un assemblage fixe de sources de lumière ponctuelles émettant une

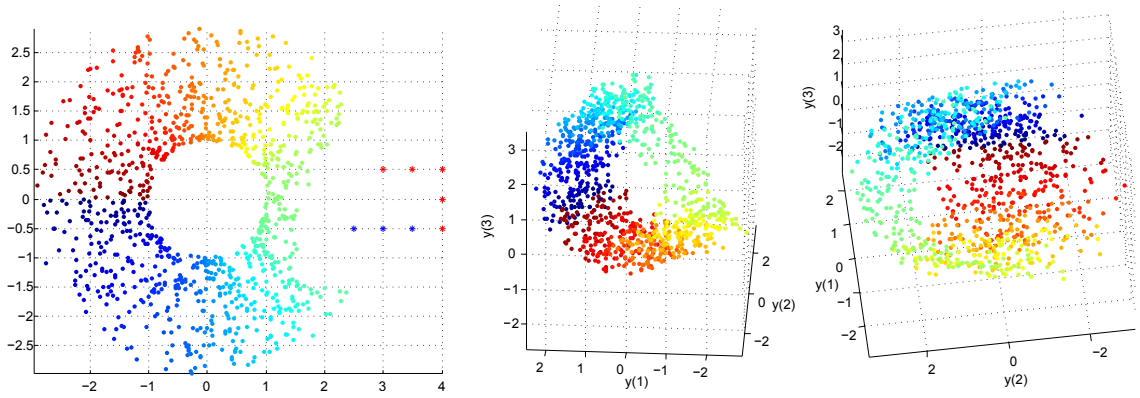


(A) Robot dans son environnement (B) Espace moteur

FIGURE 3.12 – Illustration des espaces noyaux. (Gauche) L'agent est capable d'atteindre deux configurations spatiales de capteur différentes par de multiples configurations motrices. (Droite) Celles-ci forment des espaces noyaux, prenant la forme de variétés monodimensionnelle dans cet exemple. Figure inspirée de (LAFLAQUIÈRE, 2013).

lumière rouge ou bleu. La forme de l'objet a été ici choisie de telle sorte qu'il ne présente pas de symétrie particulière : cela permet d'éviter les cas singuliers qui conduiraient à une interprétation différente entre l'agent et notre point de vue externe (ce qui, encore une fois, ne pose pas de problème du point de vue de l'agent). En tant qu'observateur externe, nous savons que la configuration spatiale de la rétine de l'agent est complètement déterminée par trois paramètres : une position (deux paramètres en translation), et une orientation (un paramètre angulaire). Ce sont bien ces 3 paramètres qui caractérisent entièrement l'interaction de l'agent avec son environnement. La question est donc : l'agent peut-il avoir accès, sur la base de son seul flux sensorimoteur, à une représentation basse-dimension de ces 3 paramètres ?

L'agent dispose donc de 4 paramètres moteurs pour fixer la configuration spatiale de la rétine dans le plan. L'espace moteur est donc redondant, et il existe un ensemble de configurations motrices qui laissent invariantes cette configuration, cf. figure 3.12a. Bien sûr, au cours de son exploration motrice naïve, l'agent ne sait pas qu'il laisse invariante la configuration spatiale de sa rétine. Par contre, il peut remarquer que différentes configurations motrices sont associées à la même valeur de sensation : c'est par ce moyen qu'il pourra détecter des invariants au sein de son flux sensorimoteur. Dans l'exemple utilisé ici, on peut remarquer que l'ensemble noyau est de dimension 1, et prend donc la forme d'une *courbe* noyau. Sous réserve de conditions a priori sur la distance relative entre la base de l'agent et l'objet, les courbes noyaux prennent la forme d'une variété monodimensionnelle bouclée dans un espace à 4 dimensions, comme illustré à la figure 3.12b. On peut y voir 2 courbes noyau (rouge et verte) $\mathcal{M}^1$ et $\mathcal{M}^2$, associées à 2 configurations spatiales différentes de la rétine. L'intuition précédente reposait sur la capacité à mesurer une distance entre ces 2 courbes noyau. Nous proposons ici d'utiliser la distance de Hausdorff (HUTTENLOCHER, KLANDERMAN et RUCKLIDGE, 1993), légèrement modifiée par rapport à sa définition initiale pour tenir compte de la nature périodique des commandes motrices utilisées. Intuitivement, cette distance peut se voir comme étant la plus grande des distances euclidiennes minimales pour se rendre des échantillons d'une courbe noyau à ceux de l'autre. Disposant ainsi d'un ensemble de distances entre de multiples paires de courbes noyau, nous pouvons alors projeter cet ensemble au sein d'un espace de dimension plus faible et visualiser



(A) Espace géométrique externe (B) Représentation interne

FIGURE 3.13 – Représentation interne obtenue par les invariants sensorimoteurs. (Gauche) Représentation des configurations spatiales du capteur de l'agent : chaque couleur est associe à une configuration angulaire particulière. (Droite) Représentation obtenue par l'agent, qui respecte la topologie de l'espace de travail. Figure tirée de (LAFLAQUIÈRE, 2013).

l'information capturée. Les simulations conduites se font de la façon suivante :

— une configuration motrice initiale $\boldsymbol{m_i}$ est choisie aléatoirement[6]. Elle est associée à la sensation $\boldsymbol{s}_i = \Psi_{\boldsymbol{\epsilon}}(\boldsymbol{m}_i)$;

— afin d'accélérer les simulations, le modèle géométrique du robot est utilisé via sa Jacobienne afin de guider l'exploration motrice sur la courbe noyau associée à la configuration motrice $\boldsymbol{m}_i$. En pratique les courbes noyaux sont discrétisées sur 100 configurations motrices ;

— on répète les 2 premières étapes jusqu'à disposer d'un échantillonnage suffisant des sensations. En pratique 1000 configurations motrices $\boldsymbol{m}_i$ aléatoires –associées à 1000 sensations $\boldsymbol{s}_i$ différentes– sont générées ;

— les distances de Hausdorff modifiées entre chacune des courbes noyau sont ensuite calculées ;

— enfin, la dimension intrinsèque de projection de ces distances est déterminée, et l'ensemble des courbes noyau est projeté via une CCA.

Le résultat de cette projection finale, obtenue dans un espace à 3 dimensions, est représenté à la figure 3.13b. Afin d'essayer d'interpréter cette représentation, l'ensemble des positions prises par le centre de la rétine au cours de l'exploration est représenté d'un point de vue externe sur la figure 3.13a selon un code couleur ne dépendant pas de son orientation. On peut constater que la représentation semble bien être une image des configurations spatiales de la rétine. Visuellement, la projection obtenue prend la forme d'un cylindre épais et creux, au sein duquel deux points proches correspondent à des points proches dans l'espace externe. Une analyse plus poussée de la représentation effectuée dans (LAFLAQUIÈRE, J. K. O'REGAN et al., 2015) montre que les positions et orientations de la rétine y sont représentées sous la forme d'une variété assimilée au produit cartésien d'un cercle (orientation de la rétine) et d'un plan (position de la rétine).

6. A nouveau, on fera tout de même en sorte que la rétine soit en capacité de visualiser l'objet présent dans l'environnement.

Publication

Le travail présenté dans cette sous-section a donné lieu à la publication de l'article (LAFLAQUIÈRE, J. K. O'REGAN et al., 2015) publié au sein de la revue "Robotics and Autonomous Systems".

3.3.1.3 Discussion

L'exemple précédent a permis d'illustrer comment la structure d'invariants sensorimoteurs permettait de capturer des caractéristiques externes à l'agent, telles que les configurations spatiales de ses capteurs. L'approche proposée détermine en quelque sorte, d'une manière non explicite, les redondances motrices de l'agent, et par là même son modèle géométrique. Cette approche purement sensorimotrice pose néanmoins quelques problèmes. Tout d'abord, la façon dont l'exploration motrice est conduite est vraisemblablement sous optimale : en cas d'exploration totalement aléatoire, il est peu probable d'échantillonner correctement les courbes noyau. Or de cet échantillonnage dépend la qualité du calcul de distance censé représenter l'interaction de l'agent avec son environnement. Il a été décidé ici de guider cet échantillonnage, via la connaissance a priori du modèle cinématique ; mais cela est uniquement possible car les ensembles noyau sont mono-dimensionnels. Pour des dimensions supérieures, le problème de l'échantillonnage reste posé. Ensuite, l'exploration motrice est conduite tandis que l'environnement ne change pas : comme précédemment, il est donc nécessaire de supposer l'environnement statique lors de la construction de la représentation interne. Pour autant, si la configuration de l'environnement change, une nouvelle représentation pourrait être construite : si elle s'avère différente dans sa forme, elle reste néanmoins équivalente (au sens mathématique) à celle obtenue pour une configuration précédente. C'est tout l'intérêt de l'approche proposée, qui est intrinsèquement indépendante de l'environnement. On pourra également noter que la représentation interne, obtenue via une projection basse dimension des distances entre courbes noyau, n'est pas strictement nécessaire sous cette forme. La projection nous permet, en tant qu'observateur externe, de vérifier que les propriétés attendues sont bien conservées. L'agent pourrait tout à fait se satisfaire de la représentation initiale, faite de paires de distances de Hausdorff. Nous en verrons d'ailleurs une exploitation dans la suite.

Pour conclure sur cette première approche naïve, peut-on dire qu'elle permet de doter l'agent de connaissances spatiales ? Nous avons en fait caractérisé les variables latentes de l'interaction, variables qui peuvent également être non spatiales (par exemple, pour un agent disposant de la capacité de contrôler sa pupille, la variable décrivant cette capacité serait incluse dans la représentation ; elle ne capture pourtant pas une caractéristique spatiale). Il faudrait donc être capable de supprimer ces descripteurs non spatiaux de la représentation. On peut espérer atteindre cet objectif en travaillant sur les capacités d'action de l'agent, c'est à dire en cherchant à travailler sur les compensations (actives) des variations sensorielles liées à un changement de configuration de l'environnement. Les travaux préliminaires proposés dans la thèse d'A. LAFLAQUIÈRE montrent que c'est envisageable. Ils ont depuis été illustrés dans (LAFLAQUIÈRE, J. O'REGAN et al., 2018), mais sans aucune formalisation mathématique là encore.

3.3.2 Vers une formalisation des ensembles noyau

Le travaux précédents effectués durant la thèse d'A. LAFLAQUIÈRE ont permis d'illustrer en quoi les ensembles noyau sont susceptibles d'être exploités par un agent pour représenter les configurations spatiales atteintes par ses capteurs. Pour autant, qu'entend-on précisément par "configurations spatiales"? L'intuition nous montre, comme illustré à la figure 3.13, qu'une information angulaire est par exemple présente au sein de la représentation interne obtenue. Mais dans un cas plus général pour lequel la dimension de cette représentation est suffisamment élevée pour être difficilement interprétable, est-il encore possible de montrer que l'agent peut avoir accès à une quelconque paramétrisation (spatiale) de son interaction? Nous touchons là du doigt les limites de l'intuition évoquée précédemment et il apparaît difficile d'imaginer aller plus loin sans tenter de formaliser plus mathématiquement l'approche. C'est précisément le but de la thèse de V. MARCEL, dont la première partie a été dédiée à la formalisation mathématique précise des ensembles noyau précédents. Ce travail est présenté dans la suite.

Nous proposons dans cette sous section de formaliser l'intuition présentée précédemment par l'introduction de considérations mathématiques qui vont nous permettre, non seulement de comprendre précisément, mais également de prouver, ce que capture la représentation interne. Pour rappel, nous faisons l'hypothèse que l'agent a seulement accès à sa configuration motrice $\boldsymbol{m} \in \mathcal{M}$, ainsi qu'à sa configuration sensorielle $\boldsymbol{s} \in \mathcal{S}$ capturant l'état physique de l'environnement. Ces deux configurations sont reliées par la loi sensorimotrice (inconnue de l'agent), de sorte que $\boldsymbol{s} = \Psi_{\boldsymbol{\epsilon}}(\boldsymbol{m})$, avec $\boldsymbol{\epsilon} \in \mathcal{E}$ la configuration de l'environnement. Il est clair que la fonction $\Psi_{\boldsymbol{\epsilon}}(.)$ est vraisemblablement non injective pour une configuration environnementale $\boldsymbol{\epsilon}$ à cause des redondances dans la géométrie de l'agent ou dans la façon dont les sensations sont générées par ses capteurs. Du point de vue de l'agent, cela veut dire que deux configurations motrice $\boldsymbol{m}_1$ et $\boldsymbol{m}_2$ peuvent, en $\boldsymbol{\epsilon}$, donner lieu à la même sensation $\boldsymbol{s} = \Psi_{\boldsymbol{\epsilon}}(\boldsymbol{m}_1) = \Psi_{\boldsymbol{\epsilon}}(\boldsymbol{m}_2)$. Une telle propriété permet de définir une relation d'équivalence $=_{\Psi_{\boldsymbol{\epsilon}}}$ entre des paires de configurations motrices, selon

$$\boldsymbol{m}_1 =_{\Psi_{\boldsymbol{\epsilon}}} \boldsymbol{m}_2 \Leftrightarrow \Psi_{\boldsymbol{\epsilon}}(\boldsymbol{m}_1) = \Psi_{\boldsymbol{\epsilon}}(\boldsymbol{m}_2). \tag{3.12}$$

Ainsi, nous pouvons regrouper toutes les configurations motrices donnant lieu à la même sensation au sein de leur classe d'équivalence $[\boldsymbol{m}]_{\boldsymbol{\epsilon}}$, définie par

$$[\boldsymbol{m}]_{\boldsymbol{\epsilon}} = \{\boldsymbol{r} \in \mathcal{M} | \boldsymbol{r} =_{\Psi_{\boldsymbol{\epsilon}}} \boldsymbol{m}\}. \tag{3.13}$$

Les classes d'équivalence peuvent alors être regroupées au sein d'un ensemble quotient moteur $\mathcal{M}/_{\boldsymbol{\epsilon}}$ donné par

$$\mathcal{M}/_{\boldsymbol{\epsilon}} = \{[\boldsymbol{m}]_{\boldsymbol{\epsilon}} \in \mathcal{P}(\mathcal{M}) | \boldsymbol{m} \in \mathcal{M}\}, \tag{3.14}$$

où $\mathcal{P}(\mathcal{M})$ désigne l'ensemble des parties de $\mathcal{M}$. Le lien entre les ensembles vus jusqu'à présent peut se résumer selon diagramme commutatif

$$\begin{array}{ccc} \mathcal{M} & \xrightarrow{\Psi_{\boldsymbol{\epsilon}}} & \mathcal{S} \\ \downarrow{\scriptstyle \pi_{\Psi_{\boldsymbol{\epsilon}}}} & \nearrow{\scriptstyle \kappa_{\boldsymbol{\epsilon}}} & \\ \mathcal{M}/_{\boldsymbol{\epsilon}} & & \end{array}, \tag{3.15}$$

où apparaît deux nouvelles applications $\pi_{\Psi_{\boldsymbol{\epsilon}}}$ et $\kappa_{\boldsymbol{\epsilon}}$. La première désigne l'application canonique surjective de $\mathcal{M}$ dans le quotient $\mathcal{M}/_{\boldsymbol{\epsilon}}$ qui à chaque élément $\boldsymbol{m}$ de $\mathcal{M}$ associe sa

classe d'équivalence $[\boldsymbol{m}]_{\boldsymbol{\epsilon}}$. La seconde est une application qui fait correspondre, par définition, chaque classe d'équivalence $[\boldsymbol{m}]_{\boldsymbol{\epsilon}}$ à une unique sensation $\boldsymbol{s}$ dans $\mathcal{S} = \Psi_{\boldsymbol{\epsilon}}(\mathcal{M})$[7]. Cette application $\kappa_{\boldsymbol{\epsilon}}$ est donc trivialement bijective.

Ce premier élément de formalisation donne déjà une indication sur ce que sont vraisemblablement les ensembles noyaux construits précédemment. Ce sont des classes d'équivalences définies par une invariance sensorielle au sein du flux sensorimoteur. Néanmoins, nous n'avons défini pour l'instant qu'un *ensemble* quotient moteur, alors que l'intuition précédente se basait sur l'existence de distances entre les différents espaces noyau.Les ensembles en question ne sont encore dotés d'aucune structure (topologique en particulier), ni même par extension de notion de distance entre les éléments les constituant. De plus, nous avons montré que l'ensemble quotient moteur était équipotent à l'ensemble des sensations. Or ces sensations sont dépendantes de l'environnement, comme illustré §3.3.1.1 : comment expliquer alors que la représentation obtenue dans la section précédente puisse être invariante à la configuration de l'environnement ? Cette question est abordée dans la suite.

3.3.2.1 L'espace des poses

D'un point de vue externe à l'agent, nous savons que les sensations $\boldsymbol{s}$ sont générées en pratique par des capteurs rigides dont l'état spatial dans le monde est entièrement décrit par leurs poses $\boldsymbol{x} \in \mathcal{X}$, avec $\mathcal{X}$ l'ensemble des poses des capteurs. Or, en tant que concepteur du système robotique, nous savons également que ces poses –habituellement exprimées en terme de coordonnés opérationnelles au sein d'un espace Euclidien– sont reliées aux configurations motrices de l'agent via la modèle géométrique direct $f(.)$, de sorte que $\boldsymbol{x} = f(\boldsymbol{m})$. Cette pose constitue donc l'état spatial du capteur, exprimé en terme de position et d'orientation par rapport à un référentiel arbitraire du monde ; elle est par contre inconnue de l'agent. Cette caractéristique extrinsèque se trouve complétée par un état sensoriel intrinsèque $\boldsymbol{s}$ correspondant à la réponse physique des transducteurs, avec $\boldsymbol{s} = \phi_{\boldsymbol{\epsilon}}(\boldsymbol{x})$, où $\phi_{\boldsymbol{\epsilon}}(.)$ désigne la fonction sensorielle directe, paramétrée par la configuration de l'environnement $\boldsymbol{\epsilon}$. Il apparaît alors que la loi sensorimotrice $\Psi_{.}(.)$ peut s'écrire comme la composition

$$\boldsymbol{s} = \Psi_{\boldsymbol{\epsilon}}(\boldsymbol{m}) = \phi_{\boldsymbol{\epsilon}}(f(\boldsymbol{m})) = (\phi_{\boldsymbol{\epsilon}} \circ f)(\boldsymbol{m}). \tag{3.16}$$

On peut remarquer que les deux fonctions f et $\phi_{\boldsymbol{\epsilon}}$. peuvent être toutes deux considérées comme surjectives. Cela est évident pour f (l'ensemble des poses des capteurs $\mathcal{X}$ est par définition l'image de $\mathcal{M}$ par f), tandis que $\phi_{\boldsymbol{\epsilon}}(.)$ peut être rendue surjective en restreignant $\mathcal{S}$ à $\mathcal{S}_{\boldsymbol{\epsilon}} = \phi_{\boldsymbol{\epsilon}}(\mathcal{X})$, avec $\mathcal{S}_{\boldsymbol{\epsilon}} \subseteq \mathcal{S}$. Cette surjectivité signifie que d'une part plusieurs configurations motrices différentes peuvent produire la même pose des capteurs (redondance de l'agent), mais d'autre part que pour une configuration de l'environnement $\boldsymbol{\epsilon}$, plusieurs poses de capteur différentes peuvent donner lieu à la même configuration sensorielle. Ce dernier cas traduit l'existence potentielle de singularités liées à la physique de l'interaction entre le capteur et son environnement, mais également d'éventuelles symétries. Il semble donc naturel de définir une nouvelle relation d'équivalence $=_{\phi_{\boldsymbol{\epsilon}}}$ telle que, pour tout $\boldsymbol{\epsilon} \in \mathcal{E}$,

$$\boldsymbol{x}_1 =_{\phi_{\boldsymbol{\epsilon}}} \boldsymbol{x}_2 \Leftrightarrow \phi_{\boldsymbol{\epsilon}}(\boldsymbol{x}_1) = \phi_{\boldsymbol{\epsilon}}(\boldsymbol{x}_2). \tag{3.17}$$

A nouveau, il est possible de regrouper toutes les poses donnant lieu à la même sensation au sein de leur classe d'équivalence $[\boldsymbol{x}]_{\boldsymbol{\epsilon}}$, définie par

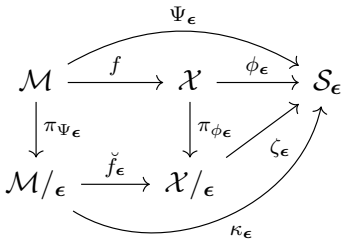
$$[\boldsymbol{x}]_{\boldsymbol{\epsilon}} = \{\boldsymbol{r} \in \mathcal{X} | \boldsymbol{r} =_{\phi_{\boldsymbol{\epsilon}}} \boldsymbol{x}\}. \tag{3.18}$$

7. Nous faisons donc l'hypothèse, pour simplifier les notations, que $\mathcal{S}$ est l'image de $\mathcal{M}$ par $\Psi_{\boldsymbol{\epsilon}}$. Si ce n'était pas le cas, il suffirait de restreindre $\mathcal{S}$ à cette image, notée $\mathcal{S}_{\boldsymbol{\epsilon}}$ dans la suite.

Et comme précédemment, ces classes d'équivalences peuvent être regroupées au sein d'un ensemble quotient des poses $\mathcal{X}/_{\epsilon}$ donné par

$$\mathcal{X}/_{\epsilon} = \{[\boldsymbol{x}]_{\epsilon} \in \mathcal{P}(\mathcal{X}) | \boldsymbol{x} \in \mathcal{X}\}. \tag{3.19}$$

L'introduction de cet espace des poses et de son quotient permet d'enrichir le diagramme commutatif précédent, qui devient maintenant



. (3.20)

Avec le même raisonnement que précédemment, nous pouvons dire que l'application ζ_{ϵ} est bijective. Et comme κ_{ϵ} l'est également, alors nous pouvons dire que l'application $\breve{f}_{\epsilon}$ est bijective. Il en ressort que les deux ensembles $\mathcal{M}/_{\epsilon}$ et $\mathcal{X}/_{\epsilon}$ sont équipotents : pour tout $\epsilon \in \mathcal{E}$, il y a autant de classe d'équivalence dans $\mathcal{M}/_{\epsilon}$ qu'il y en a dans $\mathcal{X}/_{\epsilon}$. Mais à nouveau, ces ensembles ne sont dotés d'aucune structure : la notion de distance exploitée intuitivement n'existe pas encore à ce stade. Pour autant, nous comprenons que si l'agent est capable de découvrir $\mathcal{M}/_{\epsilon}$, alors il est vraisemblablement en train de construire une représentation (interne) de $\mathcal{X}/_{\epsilon}$.

Essayons de comprendre ce que représente $\mathcal{X}/_{\epsilon}$. Au delà de sa définition mathématique (espace quotient des poses), il s'agit d'un ensemble regroupant en un même élément les poses des capteurs donnant lieu à la même sensation. Par définition, cet espace des poses quotient est donc subjectif (du moins en comparaison avec notre point de vue externe), puisque ce regroupement des poses est non seulement conditionné par les capacités perceptives de l'agent (capturées par la fonction $\phi_{.}$) mais également par la configuration ϵ de l'environnement. Ainsi, des poses de capteurs qui laisseraient invariantes des sensations rouges d'un point de vue externe pourraient ne pas être capturées dans $\mathcal{X}/_{\epsilon}$ si l'agent n'est pas doté de la capacité de voir cette couleur. De la même façon, si la configuration de l'environnement ϵ est associée à un environnement systématique associé à un ciel bleu, alors toutes les poses produisant la sensation "bleue" seront représentées par un unique élément dans $\mathcal{X}/_{\epsilon}$, alors que les configurations spatiales des capteurs sembleront elles bien distinctes d'un point de vue externe. Mais qu'adviendrait-il si pour un état $\epsilon' \neq \epsilon$ ce même ciel n'était pas identique ? L'évolution de la représentation $\mathcal{M}/_{\epsilon}$ de $\mathcal{X}/_{\epsilon}$ pour différentes configuration ϵ sera précisément formalisée dans la partie 3.3.4 de ce document, mais nous pouvons d'ores et déjà intuiter que cette représentation sera amenée à évoluer. La formalisation proposée semble donc ne pas respecter l'hypothèse posée dans la section 3.3.1 précédente, pour laquelle l'environnement était systématiquement supposé suffisamment riche pour éviter toutes les ambiguïtés listées plus haut. Cette hypothèse permettait de rendre la représentation construite sur la base des distances entre courbes noyau indépendante de ϵ. La formalisation proposée est donc ici plus générale, et nous permet de capturer plus finement les structures issues du flux sensorimoteur.

3.3.2.2 Structuration de la représentation

Comme précisé précédemment, nous n'avons prouvé l'existence que de l'équipotence entre $\mathcal{M}/_{\epsilon}$ et $\mathcal{X}/_{\epsilon}$. Essayons maintenant de préciser les éventuelles structures topologiques

pouvant montrer l'existence d'un lien plus fort entre ces deux ensembles. Les considérations mathématiques qui suivent sont principalement topologiques, et donc relativement abstraites. Pour des raisons de place nous essaierons, autant que possible, d'en fournir une interprétation intuitive, quitte à sacrifier (un peu) de justesse mathématiques ... en espérant que nous amis mathématiciens ne nous en tiendrons pas rigueur.

Sur $\mathcal{M}$: Pour commencer, nous faisons l'hypothèse que l'agent est capable de modifier ses configurations motrices $\boldsymbol{m}$ via l'application d'une succession d'actions possiblement infinitésimales. Il semble alors naturel d'imaginer que le mouvement produit est continu en temps et dans $\mathcal{M}$. Ainsi, nous faisons en fait l'hypothèse que $\mathcal{M}$ est doté d'une topologie induite par ces actions, topologie qui rend les mouvements continus. Nous allons de plus faire l'hypothèse que l'espace topologique $\mathcal{M}$ est compact[8]. Cette hypothèse implique que l'agent doit être capable d'atteindre ses bornes motrices ; cela peut être néanmoins difficile du point de vue de la commande.

Sur $\mathcal{X}$: L'ensemble des poses $\mathcal{X}$ a été introduit pour fournir un point de vue externe expliquant les configurations spatiales que pouvaient prendre les capteurs rigides de l'agent. De ce point de vue externe, ces capteurs sont capables de se déplacer dans l'espace externe, et leur déplacement est généré par des actions motrices. En particulier, une variation continue des configurations motrices correspond à un déplacement continu des capteurs dans l'espace, ce qui rend naturellement la fonction géométrique direct $f(.)$ continue. Ainsi, sans perte de généralité, on peut donner à $\mathcal{X}$ la structure topologique (la plus fine) qui rend $f(.)$ continue. Enfin, nous ferons l'hypothèse que l'espace topologique $\mathcal{X}$ est Hausdorff (ou séparé), c'est à dire que deux points distincts de $\mathcal{X}$ admettent des voisinages (i.e. des ouverts qui comprennent ces points) disjoints.

Sur les quotients $\mathcal{M}/_{\epsilon}$ et $\mathcal{X}/_{\epsilon}$: Ces deux ensembles sont deux quotients des deux espaces topologiques $\mathcal{M}$ et $\mathcal{X}$ respectivement. Ils sont donc naturellement dotés de la topologie quotient. Nous ferons également l'hypothèse que $\mathcal{X}/_{\epsilon}$ est Hausdorff. Cependant si mathématiquement cette hypothèse est fausse en général[9], on peut raisonnablement faire cette hypothèse en pratique.

De l'ensemble de ces hypothèses découle alors la propriété fondamentale suivante (relativement classique mathématiquement, mais redémontrée dans (MARCEL, ARGENTIERI et GAS, 2017)) :

$$\mathcal{M}/_{\epsilon} \text{ est homéomorphe à } \mathcal{X}/_{\epsilon}. \tag{3.21}$$

Cette propriété est bien plus forte que la simple équipotence précédente. Elle traduit le fait que les structures (topologiques) dont sont dotées $\mathcal{M}/_{\epsilon}$ (la représentation accessible à l'agent) et $\mathcal{X}/_{\epsilon}$ (celle capturée d'un point de vue externe) sont équivalentes. Elle explique ainsi en quoi travailler sur $\mathcal{M}/_{\epsilon}$ permet de capturer une structure externe à l'agent, découlant en partie des propriétés de continuité de l'action. Nous garderons en tête néanmoins que cette propriété n'est valide que pour les deux topologies quotient dont sont dotées $\mathcal{M}/_{\epsilon}$ et $\mathcal{X}/_{\epsilon}$, alors qu'une telle topologie est rarement une bonne topologie en général. Et en particulier, dès que nous chercherons à exploiter algorithmiquement ces propriétés, nous travaillerons alors avec des ensembles discrets pour lesquelles les topologies (discrètes) ne sont

8. Cette hypothèse est une généralisation de la notion d'ensemble fermé borné dans un espace Euclidien.
9. Les espaces quotient d'un espace Hausdorff ne sont pas toujours Hausdorff.

pas du tout informatives [10]. Ainsi, si (3.21) permet de capturer quelle information est représentée par $\mathcal{M}/_{\epsilon}$, il reste encore un travail théorique à mener pour comprendre comment construire, expérimentalement, une topologie qui a du sens pour l'agent. Certaines pistes sont évoquées au §3.3.4.

3.3.3 Application à la découverte du corps

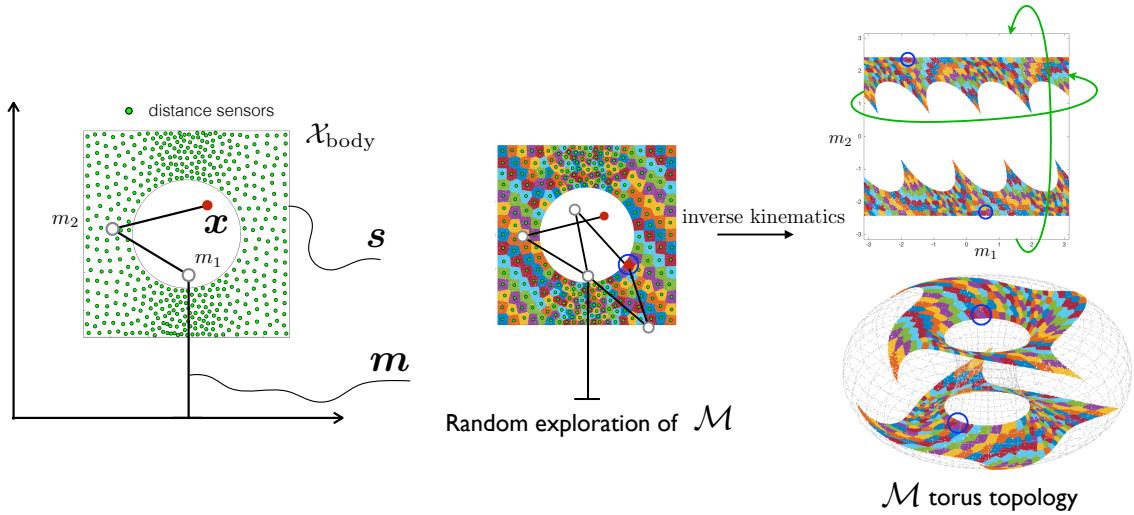
Comme la quasi totalité des travaux mentionnés précédemment, les éléments de formalisation proposés ne sont valides que pour une configuration ϵ fixe de l'environnement au cours de l'exploration motrice. Imaginer que l'environnement vérifie cette propriété constitue en soit un a priori fort. Dans un premier temps, nous avons donc cherché au début de la thèse de V. MARCEL à nous affranchir de l'influence de cet environnement en appliquant le formalisme précédent à la construction d'une représentation du corps d'un agent, ou du moins de l'interaction de celui-ci avec son propre corps. Dans cette veine, (ROSCHIN et FROLOV, 2011) a déjà proposé de traiter ensembles les informations proprioceptives et extéroceptives issues du corps en cherchant à les projeter conjointement au sein d'une représentation commune, avec l'idée que ces deux modalités doivent nécessairement partager des propriétés communes de l'espace. Ces travaux s'appuient sur un agent doté d'un bras interagissant avec son corps tactile. L'approche, bien qu'exploitant des méthodes de projections linéaires entre les deux types de modalité, permet bien de s'affranchir de l'influence de l'environnement : l'agent n'interagit qu'avec son propre corps par le toucher. Nous proposons ici d'exploiter la formalisation sensorimotrice proposée afin de construire une représentation du corps de l'agent (i) qui ne soit pas limitée par la linéarité de la technique de projection entre les modalités, et (ii) qui peut être exploitée par l'agent pour réaliser une tâche d'interpolation motrice.

3.3.3.1 Représentation sensorimotrice basse dimension du corps d'un agent

Dans toute la suite, nous allons considérer l'agent simple représenté sur la figure 3.14a. Il s'agit d'un agent plan doté d'un bras doté de 2 degrés de liberté en rotation, indépendamment pilotés par deux commandes motrices m_1 et m_2, de sorte que $\boldsymbol{m} \in \mathcal{M} =]-\pi, \pi]^2$. L'organe terminal (ponctuel) du bras est doté d'un capteur de contact informant l'agent de l'existence d'un contact avec un objet. Sa pose est totalement définie par sa position dans le plan et par son orientation, de sorte que $\mathcal{X} \in SE(2)$. L'agent est également doté d'un corps tactile, d'une forme carré, et "percé" d'un trou circulaire en son centre. Ce corps est recouvert de 300 récepteurs tactiles, générant une sensation $\boldsymbol{s} \in \mathbb{R}^{300}$ non nulle uniquement lorsque l'organe terminal du bras est en contact avec le corps. Le cas échéant, $\boldsymbol{s} = \mathbf{0}$. La sensation générée est binaire : le récepteur tactile le plus proche de l'organe terminal du bras voit sa sensation s_i fixée à 1, tandis que tous les autres récepteurs tactiles produisent en sortie une sensation nulle. Il en résulte un découpage sensoriel du corps représenté à la figure 3.14b sous la forme de patchs de couleur, indiquant quelle zone sur le corps excite quel récepteur tactile. Bien sûr, l'agent ignore tout de cette répartition.

Du point de vue externe, nous comprenons bien que la sensation $\boldsymbol{s}$ dépend uniquement de la position de l'organe terminal sur le corps, et non de son orientation. Il est donc clair que les invariants sensoriels ne sont pas des points dans $\mathcal{X} \in SE(2)$. En d'autres termes, il

10. La topologie discrète est la topologie la plus fine correspondant intuitivement au cas où chaque singleton est un ouvert de cette topologie.



(A) Corps de l'agent (B) Exploration motrice

FIGURE 3.14 – Représentation schématique de l'agent utilisé. (a) Corps de l'agent, possédant une forme carrée et creuse en son centre. (b) Représentation des clusters sensoriels sur le corps, et leur projection dans l'espace moteur.

existe plusieurs poses de l'organe terminal produisant la même sensation ; elles constitueront donc les classes d'équivalences de $\mathcal{X}/_{\epsilon} = \mathcal{X}/_{=_\phi}$ (nous gommons ainsi temporairement dans les notations la dépendance à ϵ). Le même raisonnement s'applique d'un point de vue moteur : il existe de multiples configurations motrices donnant lieu à la même sensation. De la même façon, les invariants sensoriels ne sont pas des points dans $\mathcal{M}$, et nous pouvons ainsi les regrouper au sein de classes d'équivalence dans $\mathcal{M}/_{\epsilon} = \mathcal{M}/_{=_\Psi}$. Cet aspect est illustré à la figure 3.14b : deux configurations motrices différentes stimulent le même récepteur tactile et produisent donc la même sensation. On voit bien que dans $\mathcal{M}$ ces deux configurations, entourées sur les graphiques, sont deux entités distinctes. On peut également y remarquer que grâce à la connaissance a priori du modèle géométrique du bras de l'agent, nous sommes capables (mais pas l'agent !) de projeter la position sur le corps des clusters sensoriels. L'agent de son côté n'a accès qu'aux cluster sensoriels dans son espace moteur qui peut être représenté via un espace Euclidien sur $[0, 2\pi]^2$; mais il n'en respecte pas nécessairement la topologie "naturelle", qui prend la forme d'un tore paramétré par les deux degrés de liberté en rotation du bras. La question qui se pose est donc la suivante : comment l'agent peut-il obtenir une représentation de son corps sur la base de la représentation motrice des invariants sensoriels ? Nous faisons l'hypothèse que l'agent est capable de construire la topologie quotient de $\mathcal{M}/_{=_\Psi}$ via la pseudo-métrique quotient motrice, qui consiste à dire que deux configurations motrices égales (au sens de la relation d'équivalence $=_\psi$) sont séparées d'une distance nulle. Alors, la représentation $\mathcal{M}/_{=_\Psi}$ sera bien homéomorphe à $\mathcal{X}/_{=_\phi}$, dont nous discuterons la signification un peu plus bas. Concrètement, algorithmiquement, nous procédons de la façon suivante :

— l'agent explore aléatoirement son espace des configurations motrices. Il n'en conserve que celles, notées $\boldsymbol{m}_j$, produisant une sensation non nulles $\boldsymbol{s}_j$ (i.e. pour lesquelles l'organe terminal du bras touche son corps) ;

— à chaque configuration motrice $\boldsymbol{m}_j$ correspond une sensation $\boldsymbol{s}_j$ sur le corps. En étant capable de détecter l'égalité de ces sensations [11], l'agent est capable de regrouper les

11. On suppose donc que l'agent est doté d'une "brique logique" lui indiquant quand 2 sensations sont égales.

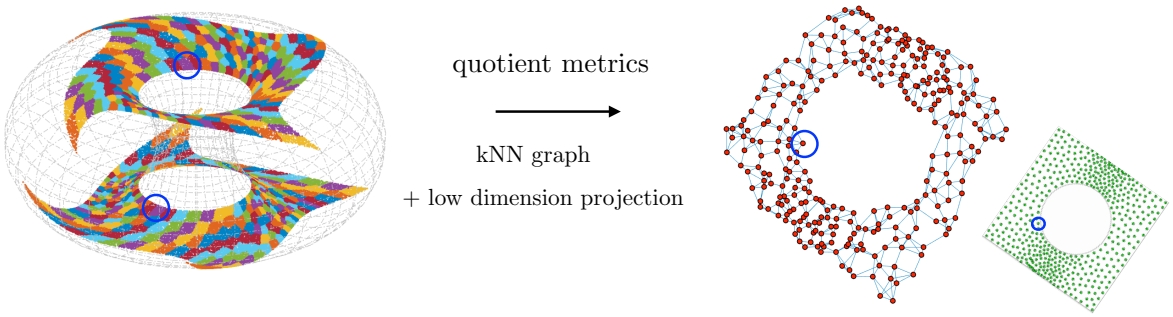


FIGURE 3.15 – Représentation basse dimension du corps de l'agent, obtenue via la structuration des invariants sensorimoteurs. La topologie de la représentation (rouge) est bien identique à celle du corps (vert). On remarque que la densité locale de la répartition des récepteurs tactiles est également bien capturée.

configurations motrices donnant lieu à la même sensation : il détermine donc les classes d'équivalence motrice $[\boldsymbol{m}_k] = \{\boldsymbol{m}_j | \boldsymbol{s}_j = \boldsymbol{s}_k\}$;

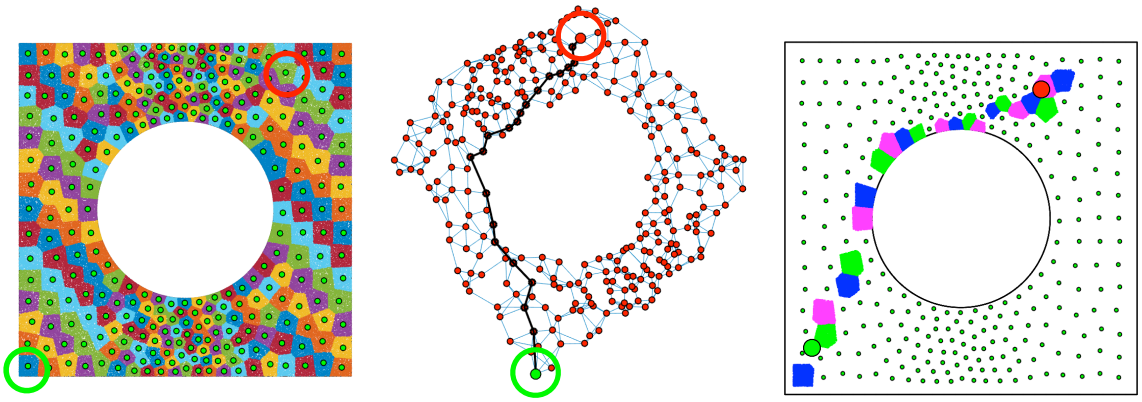
— l'agent détermine les (pseudo-)distances $\widetilde{d}$ entre les classes d'équivalence a et b formées précédemment, selon la pseudo-métrique quotient :

$$\widetilde{d}(\boldsymbol{m}_k, \boldsymbol{m}_l) = \inf\{d(\boldsymbol{p}_1, \boldsymbol{q}_1) + \ldots + d(\boldsymbol{p}_n, \boldsymbol{q}_n) | [\boldsymbol{p}_1] = \boldsymbol{m}_k, [\boldsymbol{p}_i] = [\boldsymbol{q}_{i+1}], [\boldsymbol{q}_n] = \boldsymbol{m}_l\}, \quad (3.22)$$

où $d(.,.)$ représente la distance Euclidienne entre deux configurations motrices, modifiée de telle sorte que cette distance vaut 0 lorsque deux configurations motrices appartiennent à la même classe d'équivalence ;

— A ce stade, l'agent dispose d'une représentation faite de classes d'équivalences motrices, distantes entre elles d'une pseudo-distance $\widetilde{d}(\boldsymbol{m}_k, \boldsymbol{m}_l)$. Cette représentation est donc un graphe, algorithmiquement représenté par une matrice de (pseudo-)distances entre toutes les classes d'équivalence. Un simple K-plus-proches-voisins permet de ne conserver que les voisins les plus proches, et le résultat peut être projeté en basse dimension afin d'en visualiser l'information capturée, par exemple via une CCA.

Appliquées à l'exemple illustratif proposé ici, ces différentes étapes successives produisent la représentation basse dimension de la figure 3.15. On peut y constater que le graphe obtenu est très semblable topologiquement à la disposition des récepteurs tactiles sur le corps. S'il est difficile de comparer objectivement deux topologies, nous pouvons constater que la représentation motrice des invariants sensoriels capture bien la forme du corps, de même que la densité spatiale variable de la disposition des récepteurs tactiles sur celui-ci. Sous les hypothèses posées, nous avons donc obtenu une approximation discrète de $\mathcal{M}/_{=_\Psi}$ que l'on sait homéomorphe à $\mathcal{X}/_{=_\phi}$. Dans l'exemple proposé, les sensations de l'agent sont totalement paramétrées par les positions spatiales dans le plan de l'organe terminal sur le corps. $\mathcal{X}/_{=_\phi}$ supprime donc la composante "orientation" des poses, qui n'a aucune influence sur les sensations de l'agent. Dès lors, seules ces positions sont capturées par l'agent au sein de sa représentation, qui reproduit donc la géométrie du corps. Néanmoins, il est important de noter que si la nature de l'interaction avec le corps nécessitait plus de paramètres, alors la représentation obtenue par l'agent visualisée en basse dimension ne reproduirait plus nécessairement directement une image du corps de l'agent. Cet aspect important est en particulier illustré dans (MARCEL, ARGENTIERI et GAS, 2017) pour un agent plus complexe que celui proposé ici.



(A) Corps de l'agent (B) Représentation interne (C) Chemin dans la représentation

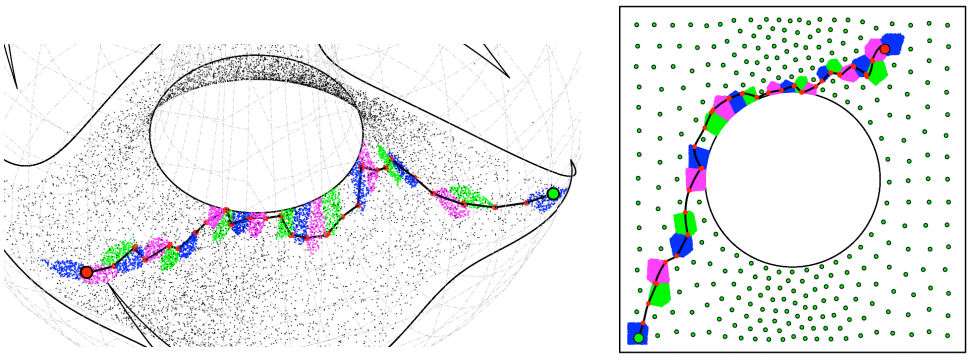
FIGURE 3.16 – Interpolation motrice au sein de la représentation. (Gauche) L'agent souhaite partir de la zone du corps entourée en vert et aller à la zone entourée en rouge. (Milieu) Les poins de départ et d'arrivée sont associés à deux classes d'équivalence, identifiés par deux nœuds dans la représentation. Une méthode de planification de chemin permet de trouver le chemin le plus court entre les deux. (Droite) Représentation sur le corps des classes d'équivalence parcourues sur le chemin.

3.3.3.2 Exploitation de la représentation : interpolation motrice

Nous n'avons pas insisté sur une autre propriété liée à l'existence d'un homéomorphisme entre $\mathcal{M}/_{=_\Psi}$ et $\mathcal{X}/_{=_\phi}$. La fonction bijective liant ces deux espace topologiques est également continue. Cela signifie que de petites variations dans $\mathcal{M}/_{=_\Psi}$ produisent de petites variations dans $\mathcal{X}/_{=_\phi}$, et réciproquement. Nous allons exploiter cette propriété de continuité afin d'effectuer une interpolation motrice au sein de la représentation. Si nous sommes capables d'y déterminer une trajectoire motrice continue, alors cela se traduira par un mouvement continu de l'organe terminal du bras de l'agent sur le corps. Cette interpolation s'effectue en 2 temps :

1. *Chemin au sein de* $\mathcal{M}/_{=_\Psi}$: il s'agit de déterminer un chemin moteur dans $\mathcal{M}/_{=_\Psi}$, de classe d'équivalence en classe d'équivalence, permettant à l'agent d'atteindre une sensation cible (sur le corps) en parcourant l'espace des poses quotient $\mathcal{X}/_{=_\phi}$ de manière continue. L'approximation de $\mathcal{M}/_{=_\Psi}$ étant faite d'un graphe constitué des distances entre classes d'équivalence les plus proches, il s'agit donc de planifier un chemin dans un graphe, partant d'un nœud (une classe d'équivalence de départ, choisie éventuellement au hasard selon la tâche) pour arriver au nœud final (la classe d'équivalence associée la sensation cible), par exemple via l'algorithme de Dijkstra (DIJKSTRA, 1959) ou d'autres approches (BONDY et MURTY, 2007). Cette première étape est illustrée à la figure 3.16.
2. *Trajectoire dans l'espace* $\mathcal{M}$: le bras de l'agent est concrètement piloté via les commandes motrices m_1 et m_2. Il s'agit donc d'en déterminer les valeurs successives permettant de passer d'une classe d'équivalence à une autre (en respectant le chemin déterminé à l'étape 1), mais également pour traverser une classe d'équivalence. Nous avons proposé dans (MARCEL, ARGENTIERI et GAS, 2017) une approche itérative cherchant à identifier les configurations motrices intermédiaires minimisant le trajet entre et à l'intérieur des classes d'équivalence. Une fois ces configurations déterminées, un algorithme standard d'interpolation linéaire est utilisé pour trouver les configurations intermédiaires. Cette étape est illustrée à la figure 3.17.

Une fois le chemin moteur déterminé, son exécution par l'agent conduit à un mouvement fluide et continu de l'organe terminal du bras sur le corps de l'agent. Ainsi, sur la base d'une



(A) Interpolation dans l'espace moteur

(B) Mouvement correspondant sur le corps

FIGURE 3.17 – Interpolation motrice dans l'espace $\mathcal{M}$. (Gauche) Une fois le chemin moteur balisé par les classes d'équivalence (représentées sous forme de patchs de couleur), il s'agit de déterminer le chemin moteur pour les rejoindre et les traverser. (Droite) Une fois les configurations motrices déterminées, leur application produit un mouvement continu de l'organe terminal du bras sur le corps.

représentation des invariants sensorimoteurs, l'agent est capable de produire, d'un point de vue externe, un mouvement continu de son bras pour atteindre n'importe quelle sensation tactile, et donc n'importe quelle partie de son corps. A nouveau, une illustration dans un cas plus complexe est proposée dans (MARCEL, ARGENTIERI et GAS, 2017). L'interpolation motrice y est effectuée sur une représentation de plus grande dimension, ne capturant pas visuellement la géométrie du corps. Pour autant, les mouvements générés par le bras font toujours en sorte de parcourir celui-ci d'une manière continue. Cet exemple supplémentaire illustre à son tour la pertinence de la représentation obtenue, et la nature des déplacements opérés est garantie par la continuité de cette représentation.

Publication

L'ensemble de ces travaux, de la formalisation générale des invariants sensorimoteurs en passant par son application à la découverte du corps d'un agent et l'interpolation motrice ont donné lieu à l'article (MARCEL, ARGENTIERI et GAS, 2017) publié au sein de la revue IEEE "Transactions on Cognitive and Developmental Systems".

3.3.4 Raffinement de la représentation tout au long de la vie de l'agent

Les travaux précédents de V. MARCEL ont permis de montrer comment exploiter le formalisme des ensembles quotient pour la construction d'une représentation de l'interaction qu'a l'agent avec son propre corps. Pour cela, nous avons temporairement mis de côté la dépendance explicite dans les équations de la représentation à l'environnement en faisant en sorte que l'expérience sensorimotrice soit restreinte au corps de l'agent uniquement. C'est bien entendu une hypothèse forte et nous avons souhaité attaquer au cours de la seconde partie de thèse de V. MARCEL le problème de cette dépendance à l'environnement, en imaginant comment un agent pourrait, au fur et à mesure de son interaction avec différentes configurations de l'environnement, modifier la représentation de son interaction. Ce travail, encore préliminaire, est présenté en partie dans la suite.

3.3.4.1 Éléments de formalisation

Revenons pour cela à la définition de la relation d'équivalence (3.12), rappelée ici :

$$\boldsymbol{m}_1 =_{\Psi_{\boldsymbol{\epsilon}}} \boldsymbol{m}_2 \Leftrightarrow \Psi_{\boldsymbol{\epsilon}}(\boldsymbol{m}_1) = \Psi_{\boldsymbol{\epsilon}}(\boldsymbol{m}_2). \tag{3.23}$$

Pour rappel également, les configurations motrices donnant lieu aux même sensation sont ainsi regroupées au sein de leur classe d'équivalence $[\boldsymbol{m}]_{\boldsymbol{\epsilon}} = \{\boldsymbol{r} \in \mathcal{M} | \boldsymbol{r} =_{\Psi_{\boldsymbol{\epsilon}}} \boldsymbol{m}\}$. Il est connu que l'ensemble des classes d'équivalence forment une partition [12] de l'ensemble sur lequel la relation d'équivalence est définie. En d'autres termes, chaque élément de $\mathcal{M}$ appartient à une seule et unique classe d'équivalence $[\boldsymbol{m}]_{\boldsymbol{\epsilon}}$. Ainsi, l'ensemble quotient $\mathcal{M}/_{\boldsymbol{\epsilon}}$ forme un raffinement de la partition triviale $\{\mathcal{M}\}$. Cette propriété de raffinement pour s'écrire formellement

$$\mathcal{M}/_{\boldsymbol{\epsilon}} \leq \{\mathcal{M}\}, \tag{3.24}$$

où $\leq$ désigne la relation d'ordre partielle "plus fin que". On dit d'une partition X qu'elle est plus fine qu'une partition Y si chaque élément dans X est inclus dans un élément de Y : X est alors composé de parties fragmentées de Y. On voit alors que la relation d'ordre (3.24) peut être complétée par

$$\{\{\boldsymbol{m}\} | \boldsymbol{m} \in \mathcal{M}\} \leq \mathcal{M}/_{\boldsymbol{\epsilon}} \leq \{\mathcal{M}\}. \tag{3.25}$$

Ainsi, la partition la plus fine est celle composée des singletons $\{\boldsymbol{m}\}$, composée individuellement d'une seule et unique configuration motrice. Au contraire, la partition $\{\mathcal{M}\}$ est la partition la plus grossière de $\mathcal{M}$. A ce titre, et par abus de notation, nous pourrions d'ailleurs écrire que $\{\mathcal{M}\} = \mathcal{M}/_{\emptyset}$: la partition motrice la plus grossière est obtenue pour un agent n'ayant encore jamais interagi avec un environnement.

Considérons maintenant que la configuration de l'environnement change, pour passer de $\boldsymbol{\epsilon}$ à $\boldsymbol{\epsilon}'$. Si l'agent a connaissance de ce changement, il peut alors utiliser la nouvelle relation d'équivalence $=_{\psi_{\boldsymbol{\epsilon}'}}$, donnant lieu à une nouvelle partition motrice $\mathcal{M}/_{\boldsymbol{\epsilon}'}$. Notons bien que pour n'importe quel $\boldsymbol{m}_1, \boldsymbol{m}_2 \in \mathcal{M}$, $\boldsymbol{m}_1 =_{\Psi_{\boldsymbol{\epsilon}}} \boldsymbol{m}_2 \not\Leftrightarrow \boldsymbol{m}_1 =_{\Psi_{\boldsymbol{\epsilon}'}} \boldsymbol{m}_2$, ou encore $\boldsymbol{m}_1 \neq_{\Psi_{\boldsymbol{\epsilon}}} \boldsymbol{m}_2 \not\Leftrightarrow \boldsymbol{m}_1 \neq_{\Psi_{\boldsymbol{\epsilon}'}} \boldsymbol{m}_2$. En fait, les deux partitions motrices ne peuvent pas être comparées directement, car issues du partitionnement de $\mathcal{M}$ pour deux configurations différentes de l'environnement. Cependant, nous pouvons définir une nouvelle relation d'équivalence *multi-environnement* $=_{\Psi_{(\boldsymbol{\epsilon},\boldsymbol{\epsilon}')}}$ selon

$$\boldsymbol{m}_1 =_{\Psi_{(\boldsymbol{\epsilon},\boldsymbol{\epsilon}')}} \boldsymbol{m}_2 \Leftrightarrow \begin{cases} \boldsymbol{m}_1 =_{\Psi_{\boldsymbol{\epsilon}}} \boldsymbol{m}_2 \\ \quad \text{et} \\ \boldsymbol{m}_1 =_{\Psi_{\boldsymbol{\epsilon}'}} \boldsymbol{m}_2 \end{cases}. \tag{3.26}$$

Dès lors, cette nouvelle relation d'équivalence conduit à de nouvelles classes d'équivalence $[\boldsymbol{m}]_{(\boldsymbol{\epsilon},\boldsymbol{\epsilon}')}$ vérifiant $[\boldsymbol{m}]_{(\boldsymbol{\epsilon},\boldsymbol{\epsilon}')} \subseteq [\boldsymbol{m}]_{\boldsymbol{\epsilon}}$ et $[\boldsymbol{m}]_{(\boldsymbol{\epsilon},\boldsymbol{\epsilon}')} \subseteq [\boldsymbol{m}]_{\boldsymbol{\epsilon}'}$. En terme de relation d'ordre, cela se traduit par

$$\begin{cases} \{\{\boldsymbol{m}\} | \boldsymbol{m} \in \mathcal{M}\} \leq \mathcal{M}/_{(\boldsymbol{\epsilon},\boldsymbol{\epsilon}')} \leq \mathcal{M}/_{\boldsymbol{\epsilon}} \leq \{\mathcal{M}\} \\ \quad \text{et} \\ \{\{\boldsymbol{m}\} | \boldsymbol{m} \in \mathcal{M}\} \leq \mathcal{M}/_{(\boldsymbol{\epsilon},\boldsymbol{\epsilon}')} \leq \mathcal{M}/_{\boldsymbol{\epsilon}'} \leq \{\mathcal{M}\} \end{cases}, \tag{3.27}$$

ce qui montre que par définition, $\mathcal{M}/_{(\boldsymbol{\epsilon},\boldsymbol{\epsilon}')}$ est un raffinement des deux partitions $\mathcal{M}/_{\boldsymbol{\epsilon}}$ et $\mathcal{M}/_{\boldsymbol{\epsilon}'}$.

On peut remarquer que la nouvelle relation d'équivalence $=_{\Psi_{(\boldsymbol{\epsilon},\boldsymbol{\epsilon}')}}$ ne dépend pas de l'ordre dans lequel les deux configurations $\boldsymbol{\epsilon}$ et $\boldsymbol{\epsilon}'$ ont été expérimentées par l'agent. Par

12. On rappelle qu'une partition d'un ensemble X est un ensemble de sous ensembles non vides, deux à deux disjoints, et dont l'union forme l'ensemble X.

conséquent, la paire (ϵ, ϵ') peut être écrite sous la forme d'un sous ensemble $E = \{\boldsymbol{\epsilon}, \boldsymbol{\epsilon}'\}$. Ainsi, nous pouvons généraliser la relation d'équivalence (3.26) en une nouvelle relation d'équivalence générique multi-environnement $=_{\Psi_E}$, pour tout $E \in \mathcal{E}$, selon

$$\begin{aligned} \boldsymbol{m}_1 =_{\Psi_E} \boldsymbol{m}_2 &\Leftrightarrow \forall \boldsymbol{\epsilon} \in E, \Psi_{\boldsymbol{\epsilon}}(\boldsymbol{m}_1) = \Psi_{\boldsymbol{\epsilon}}(\boldsymbol{m}_2), \\ &\Leftrightarrow \Psi_E(\boldsymbol{m}_1) = \Psi_E(\boldsymbol{m}_2), \end{aligned} \tag{3.28}$$

où la fonction Ψ_E fait correspondre chaque configuration motrice $\boldsymbol{m}$ à sa *séquence sensorielle* acquise tout au long des changement successifs de configurations de l'environnement $\boldsymbol{\epsilon} \in E$, et donc acquise tout au long de la vie de l'agent, avec

$$\begin{aligned} \Psi_E : \quad \mathcal{M} &\to \prod_{\boldsymbol{\epsilon} \in E} \mathcal{S} \\ \boldsymbol{m} &\mapsto (\Psi_{\boldsymbol{\epsilon}}(\boldsymbol{m}))_{\boldsymbol{\epsilon} \in E} \end{aligned} \tag{3.29}$$

où $\prod$ désigne le produit Cartésien des ensembles. On peut alors déduire certaines propriétés de cette mise en équation. Pour commencer, considérons deux sous-ensembles non vides $E_1, E_2 \subseteq \mathcal{E}$ tels que $E_2 \subseteq E_1$. Nous avons alors

$$\mathcal{M}/_{E_1} \leq \mathcal{M}/_{E_2}, \tag{3.30}$$

traduisant le fait que la partition $\mathcal{M}/_{E_1}$ est plus fine que $\mathcal{M}/_{E_2}$. Dès lors, on comprend que le cas où $E = \mathcal{E}$ conduit à la partition la plus fine à laquelle l'agent peut avoir accès. On peut alors écrire

$$\{\{\boldsymbol{m}\} | \boldsymbol{m} \in \mathcal{M}\} \leq \mathcal{M}/_{\mathcal{E}} \leq \mathcal{M}/_{E} \leq \{\mathcal{M}\}, \forall E \in \mathcal{P}(\mathcal{E}). \tag{3.31}$$

Au sein de cette équation, $\mathcal{M}/_{\mathcal{E}}$ est particulièrement important. Cette partition est la plus fine accessible à l'agent. Elle est donc constituée de classes d'équivalence qui ne pourront jamais plus être fragmentées au cours de l'expérience sensorimotrice. En ce sens, ces classes d'équivalence constituent ce que nous appellerons les *points sensorimoteurs les plus fins*, en référence à la notion de points bien connue a priori d'un point de vue externe.

Sur la base du même raisonnement, nous pouvons introduire –comme nous l'avons fait dans la formalisation précédente– l'espace des poses $\mathcal{X}$ caractérisant les poses $\boldsymbol{x} \in \mathcal{X}$ prises dans l'espace par les capteurs de l'agent d'un point de vue externe. Nous avions pour cela introduit la relation d'équivalence (3.17), qui nous permet immédiatement, selon la même approche que (3.28), de définir la nouvelle relation d'équivalence multi-environnement $=_{\phi_E}$ selon

$$\begin{aligned} \boldsymbol{x}_1 =_{\phi_E} \boldsymbol{x}_2 &\Leftrightarrow \forall \boldsymbol{\epsilon} \in E, \phi_{\boldsymbol{\epsilon}}(\boldsymbol{x}_1) = \phi_{\boldsymbol{\epsilon}}(\boldsymbol{x}_2), \\ &\Leftrightarrow \phi_E(\boldsymbol{x}_1) = \phi_E(\boldsymbol{x}_2), \end{aligned} \tag{3.32}$$

où la fonction ϕ_E fait correspondre chaque pose $\boldsymbol{x}$ à sa séquence sensorielle acquise tout au long de la vie de l'agent, avec

$$\begin{aligned} \phi_E : \quad \mathcal{X} &\to \prod_{\boldsymbol{\epsilon} \in E} \mathcal{S} \\ \boldsymbol{x} &\mapsto (\phi_{\boldsymbol{\epsilon}}(\boldsymbol{x}))_{\boldsymbol{\epsilon} \in E} \end{aligned} \tag{3.33}$$

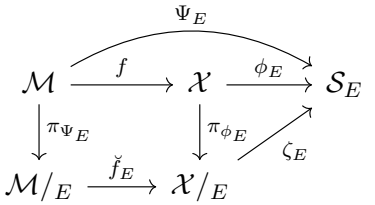
La relation d'équivalence $=_{\phi_E}$ peut ainsi être comprise comme : deux poses sont considérées égales après avoir vu tous les environnements dans $E \subseteq \mathcal{E}$ si leurs sensations associées sont égales pour tous les environnements dans E. Nous pouvons donc à nouveau regrouper ces poses au sein de leur classe d'équivalence $[\boldsymbol{x}]_E$, qui définissent à leur tour un ensemble

quotient $\mathcal{X}/_E$ se trouvant être une partition plus fine que $\mathcal{X}/_\epsilon$. Il vient alors de manière immédiate

$$\{\{\boldsymbol{x}\}|\boldsymbol{x} \in \mathcal{X}\} \leq \mathcal{X}/_\mathcal{E} \leq \mathcal{X}/_E \leq \{\mathcal{X}\}, \forall E \in \mathcal{P}(\mathcal{E}). \tag{3.34}$$

Comme précédemment, le cas $E = \mathcal{E}$ est particulièrement intéressant. En effet, $\mathcal{X}/_\mathcal{E}$ est constitué de classes d'équivalences qui ne peuvent pas être fragmentées plus, et définit ainsi *l'espace des poses sensibles le plus fin*. D'un point de vue externe, cela traduit le fait que certaines positions/orientations des capteurs ne peuvent intrinsèquement pas être distinguées l'une de l'autre par leurs valeurs de sensations associées. L'agent ne sera donc pas capable, sur la base de son seul flux sensorimoteur, de construire une représentation capable de séparer ces poses : nous venons donc de définir une borne minimale atteignable par l'agent, dépendant de ses capacités sensorimotrices.

Le lien entre tous les ensembles définis précédemment peut se résumer en le diagramme commutatif suivant, très proche de (3.20)

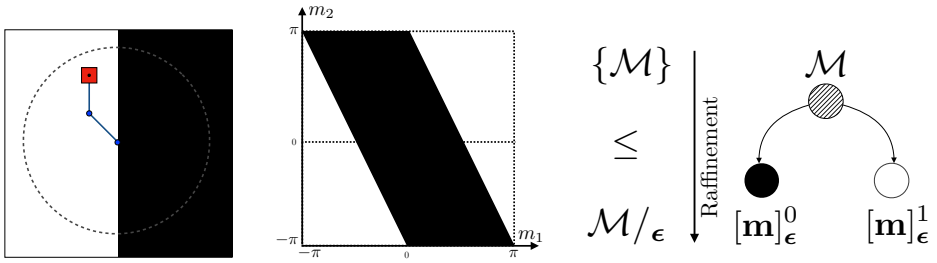


. (3.35)

Ainsi, de manière évidente, nous avons à nouveau la propriété d'équipotence entre $\mathcal{M}/_E$ et $\mathcal{X}/_E$. Et nous faisons donc l'hypothèse que $\mathcal{M}/_E$ peut être utilisé comme représentation interne de $\mathcal{X}/_E$. Et en particulier, dans le cas où $E = \mathcal{E}$, à chaque point sensorimoteur le plus fin (dans $\mathcal{M}/_\mathcal{E}$) correspond une unique pose sensible la plus fine (dans $\mathcal{X}/_\mathcal{E}$). A nouveau, nous n'avons pour l'instant eu aucune considération topologique. Néanmoins, comme le laisse présager l'exemple illustratif qui suit, une structure plus forte, comme l'homéomorphisme, semble préserver les structures dans les espaces quotient considérés. Ce travail – encore en cours– porte sur l'introduction de métriques probabilistes couplées à des notions d'observabilité et permet d'introduire des topologies naturelles démontrant l'existence d'un homéomorphisme entre ces espaces.

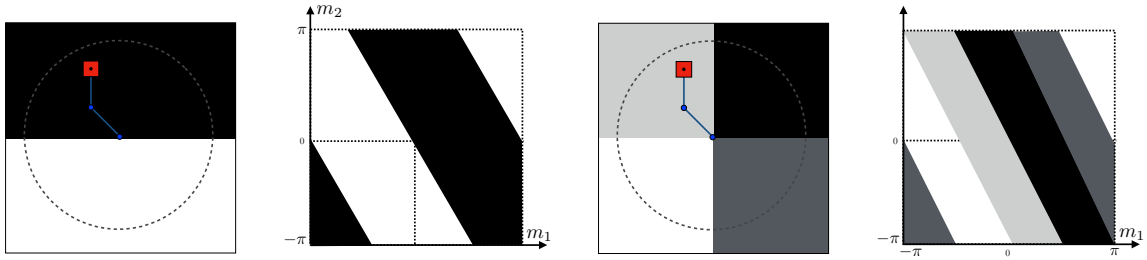
3.3.4.2 Illustration du raffinement

Afin d'illustrer le principe du raffinement, nous proposons d'utiliser un agent plan doté d'un bras série à deux degrés de libertés en rotation, commandés par deux commandes motrices m_1 et m_2 tels que $m_1, m_2 \in [-\pi; \pi[$ (par convention $m_1 = m_2 = 0$ rend le bras horizontal). Au bout du bras est placé une caméra ponctuelle uniquement sensible à l'illumination, et générant un scalaire $\boldsymbol{s} = s = 0$ si l'illumination est nulle, et $s = 1$ sinon. Le bras, fixe dans son environnement, conduit la caméra à explorer son environnement selon un disque centré sur sa base. L'environnement est supposé constitué d'une zone blanche et d'une zone noire, séparées toutes deux par une droite, cf. figure 3.18a. Au début de la vie de l'agent, celui-ni ne dispose d'aucune connaissance a priori. L'ensemble des configurations motrices n'a pas encore été raffiné, de sorte que $\{\mathcal{M}\}$ est la partition la plus fine dont il dispose. Après une première exploration de l'environnement noir et blanc, l'agent est capable de former la partition $\mathcal{M}/_\epsilon = \{[\boldsymbol{m}]^0_\epsilon, [\boldsymbol{m}]^1_\epsilon\}$, qui peut être représentée directement dans son espace des configurations motrices (figure 3.18b), ou sous la forme d'un graphe à deux nœuds (figure 3.18c). A ce moment de la vie de l'agent, celui-ci n'a fait l'expérience que de la configuration ϵ de l'environnement. En ce sens, $\mathcal{M}/_\epsilon$ définit alors la représentation la plus



(A) Environnement ϵ (B) Clusters sensorimoteurs pour ϵ (C) Illustration du raffinement

FIGURE 3.18 – Illustration du raffinement. L'environnement délimite deux clusters sensorimoteurs, associés aux sensations $s = 0$ ou $s = 1$. La partition motrice initiale peut alors être raffinée en deux sous ensembles, les deux classes d'équivalence $[\boldsymbol{m}]^0_{\epsilon}$ et $[\boldsymbol{m}]^1_{\epsilon}$ respectivement.



(A) Environnement ϵ' (B) Clusters sensorimoteurs pour ϵ' (C) Illustration de la séquence (ϵ, ϵ') (D) Clusters sensorimoteurs pour (ϵ, ϵ')

FIGURE 3.19 – Illustration du raffinement (suite). Une nouvelle configuration ϵ' de l'environnement est présentées à l'agent, donnant lieu à de nouveaux clusters dans $\mathcal{M}$, difficilement comparables à ceux obtenus pour ϵ. La prise en compte de la séquence (ϵ, ϵ') permet de visualiser le raffinement de la partition obtenue avec ϵ seul.

fine accessible à l'agent. Si jamais l'état de l'environnement ne changeait plus, l'agent ne serait pas capable de raffiner sa représentation qui se retrouverait alors réduite à deux points sensorimoteurs.

Considérons maintenant que la configuration de l'environnement change de ϵ pour aller en ϵ'. Cette nouvelle configuration correspond à une nouvelle séparation de l'espace de travail en deux zones blanches et noires, cf. figure 3.19a. Comme formalisé précédemment, les deux partitions obtenues pour ϵ et ϵ' ne sont pas directement comparables. C'est la prise en compte *simultanée* des deux configurations de l'environnement qui permet alors de comprendre comment l'agent peut affiner sa représentation obtenue avec ϵ. On peut alors constater sur la figure 3.19d que cette séquence permet de scinder la classe d'équivalence $[\boldsymbol{m}]^0_{\epsilon}$ (représentée en noir sur la figure 3.18b) en deux nouvelles classes d'équivalence $[\boldsymbol{m}]^{00}_E$ et $[\boldsymbol{m}]^{01}_E$ (représentées et noir et gris foncé sur la figure 3.19d), avec $E = \{\epsilon, \epsilon'\}$. Le même raisonnement s'applique pour la classe d'équivalence $[\boldsymbol{m}]^1_{\epsilon}$ (représentée en blanc sur la figure 3.18b) qui se fragmente en 2 sous ensembles $[\boldsymbol{m}]^{10}_E$ et $[\boldsymbol{m}]^{11}_E$ (représentées en gris clair et en blanc sur la figure 3.19d). Ce raffinement de la partition motrice peut à nouveau se représenter sous la forme d'un graphe, tel que celui représenté sur la figure 3.20. Dans l'exemple utilisé, on peut aisément imaginer une troisième configuration de l'environnement qui conduirait au fractionnement de certaines classes d'équivalence. Ainsi, de manière itérative, il est clair que le nombre de nœuds du graphe ira en augmentant et tendra même vers l'infini dans ce cas précis : il y aura toujours une nouvelle configuration de l'environnement (i.e. une séparation en noir et blanc de l'espace de travail de l'agent) qui conduira à un nouveau fractionnement d'une ou plusieurs classe(s) d'équivalence.

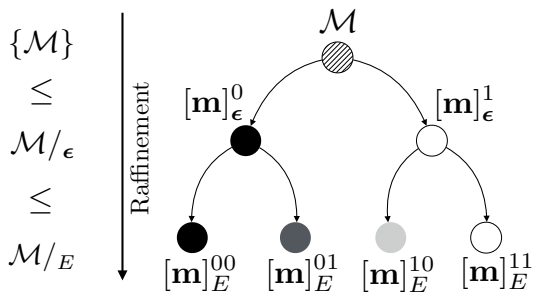


FIGURE 3.20 – L'agent, en faisant l'expérience successive de deux états différents de l'environnement, est capable fractionner successivement sa représentation motrice de son interaction avec l'environnement. Cette opération conduit au raffinement de la représentation, au fur et à mesure de la vie de l'agent.

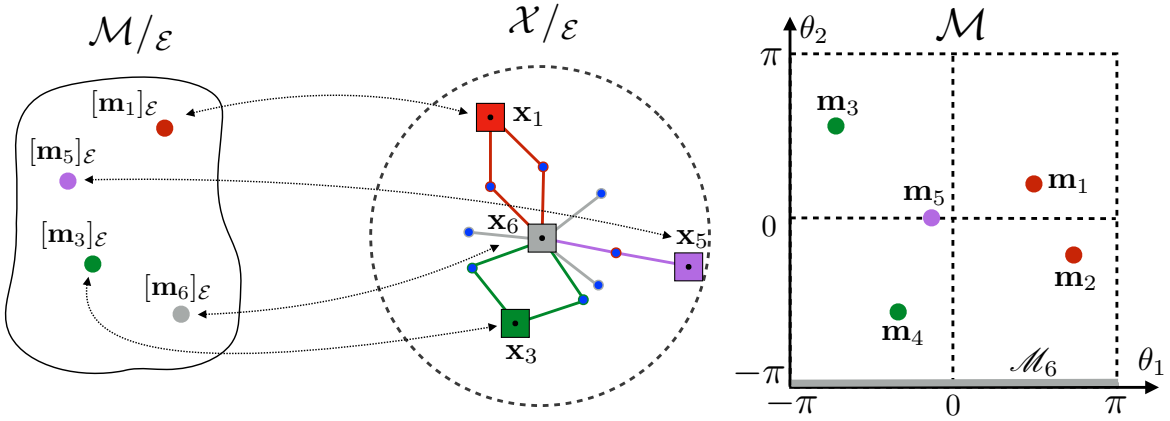
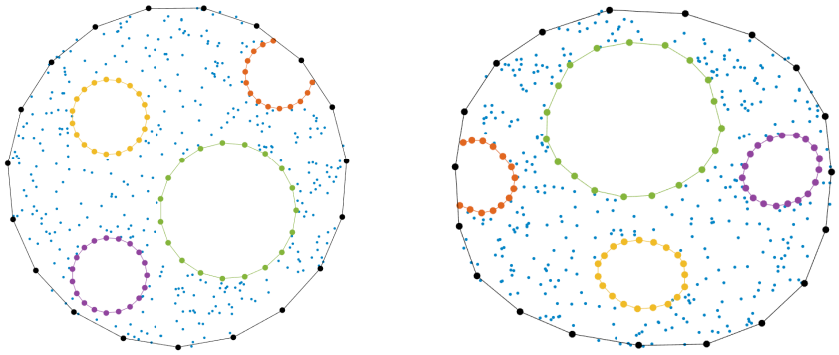


FIGURE 3.21 – Illustration de la signification du raffinement le plus fin. Aux points sensorimoteurs les plus fins correspondent un point dans l'espace des poses sensibles le plus fin. De par l'interaction ponctuelle de l'agent avec son environnement, ces points dans l'espace des poses sensibles le plus fin sont en fait identiques aux point 2D dans l'espace de travail de l'agent.

Pour l'instant, nous n'avons pris que le point de vue de l'agent, qui a donc été capable d'affiner sa représentation grâce à son interaction avec deux configurations différentes de l'environnement. Mais que capture cette représentation, en particulier dans sa version *la plus fine* $\mathcal{M}/_\mathcal{E}$ pour cet exemple ? Si on décide d'adopter un point de externe à l'agent (c'est à dire un point de vue qui nous dote de toutes les connaissances a priori non accessibles à l'agent), nous savons que celui-ci est doté d'un capteur ponctuel : une pose $\boldsymbol{x}$ n'est donc paramétrée que par les deux positions x et y de ce point dans le plan. Et comme nous savons que deux points dans un espace Euclidien 2D peuvent toujours être séparés par une droite, il existera toujours une configuration $\boldsymbol{\epsilon}$ de l'environnement pour laquelle les sensations en ces deux points seront distinctes. On en déduit alors que $[\boldsymbol{x}]_\mathcal{E} = \{\boldsymbol{x}\}$: la partition la plus fine de l'espace de travail de l'agent est données par $\mathcal{X}/_\mathcal{E} = \{\{(x, y\}, (x, y) \in \mathcal{X}\}$. Nous sommes en fait dans un cas particulier où les classes d'équivalence de l'ensemble des poses sensibles le plus fin sont exactement les points dans l'espace des poses, et donc ici les points dans l'espace de travail. Ainsi, de part le lien d'équipotence établi entre $\mathcal{M}/_\mathcal{E}$ et $\mathcal{X}/_\mathcal{E}$, nous pouvons dire qu'à chacun des points sensorimoteurs les plus fins (i.e. les classes d'équivalences découvertes par l'agent dans $\mathcal{M}/_\mathcal{E}$) correspond un point dans l'espace des poses sensibles le plus fin, et donc correspond également un point dans l'espace de travail. Cette constatation est illustrée à la figure 3.21, où les trois ensembles $\mathcal{M}$, $\mathcal{M}/_\mathcal{E}$ et $\mathcal{X}/_\mathcal{E}$ sont schématiquement représentés. L'espace des poses sensibles le plus fin est représenté au centre : chacune des



(A) Représentation de l'espace de travail accessible à l'agent.

(B) Représentation interne obtenue par l'agent.

FIGURE 3.22 – Conservation de la topologie de l'espace de travail au sein de la représentation interne $\mathcal{M}/_E$, avec E suffisamment grand.

poses $\boldsymbol{x}_i$ est représentée par une position 2D dans l'espace, atteinte possiblement par un ensemble de configurations motrices. Dans cette illustration, on peut par exemple voir que les deux configurations motrices $\boldsymbol{m}_1$ et $\boldsymbol{m}_2$ permettent toutes deux d'atteindre la pose $\boldsymbol{x}_1$. Dans la représentation sensorimotrice la plus fine, l'agent fait donc correspondre à cette pose un unique point sensorimoteur $[\boldsymbol{m}_1]_\mathcal{E}$. La même chose s'applique pour la pose $\boldsymbol{x}_6$, atteinte par l'ensemble des configuration motrice $\mathscr{M}_6$, à son tour représentée par un unique point $[\boldsymbol{m}_6]_\mathcal{E}$. L'agent a donc a priori accès à une représentation de son espace de travail sur la seule base de l'analyse des invariants de son flux sensorimoteur. D'une certaine manière, on comprend qualitativement que l'application allant de $\mathcal{M}$ vers $\mathcal{M}/_\mathcal{E}$ converge vers une application possédant le même noyau que la fonction géométrique directe f : l'agent a capturé l'intégralité de sa redondance géométrique.

La figure 3.21 illustre un lien point à point entre les $\mathcal{M}/_\mathcal{E}$ et $\mathcal{X}/_\mathcal{E}$. Mais pour être une bonne représentation, il faudrait également que $\mathcal{M}/_\mathcal{E}$ capture la même topologie que $\mathcal{X}/_\mathcal{E}$. Pourtant, on peut pressentir qu'un tel lien existe. Par exemple, le graphe obtenu sur la figure 3.20 est un graphe orienté, mais non connecté : à une étape du raffinement, il ne semble pas y avoir de lien entre les différentes classes d'équivalence. Cependant, elles sont issues de classes d'équivalence communes à l'étape précédente. Ainsi, les ensembles $[\boldsymbol{m}]^{00}_E$ et $[\boldsymbol{m}]^{01}_E$ (noir et gris foncés respectivement) partagent le même "ancêtre" $[\boldsymbol{m}]^0_\epsilon$. Et on constate au sein de la figure 3.19c que ces classes d'équivalence sont associées à deux zones spatiales (colorées de manière identique) de l'espace de travail proches, séparées par une frontière commune. On sent bien que des propriétés topologiques, dont l'existence n'a pas encore été prouvée au sein du formalisme précédent, permettent de capturer que des zones proches spatialement seront associées à des classes d'équivalences proches dans le partitionnement de $\mathcal{M}$ opéré par l'agent. Au sein de la figure 3.21, cela se traduirait par des classes d'équivalence $[\boldsymbol{m}_i]_\mathcal{E}$ d'autant plus proches que leurs poses associées $\boldsymbol{x}_i$ sont proches dans l'espace. Sans en détailler plus les paramètres de simulation, des résultats préliminaires illustrent que cette propriété est effectivement vérifiée. La figure 3.22 représente l'espace de travail exploré par l'agent (doté du même type de capacités motrices et sensorielles que précédemment), volontairement parsemé de zones inaccessibles dont les cercles de couleurs représentent les frontières. La représentation interne $\mathcal{M}/_E$ obtenue pour un ensemble E suffisamment grand (ce qui peut être vu comme "une vie suffisamment longue" de l'agent), projetée en basse dimension, conserve bien la topologie de l'espace de travail. Bien sûr, dans des cas où l'interaction de l'agent avec son environnement ne se résume pas à une physique ponctuelle, l'interprétation directe de la représentation peut être plus complexe pour nous, observateurs

externes. Néanmoins, du point de vue de l'agent, elle contient tous les variables latentes caractérisant son interaction avec l'environnement.

Publication

La contribution présentée dans cette sous-section est adaptée d'une partie d'un article en cours de rédaction par V. MARCEL.

Chapitre 4

Conclusion

4.1 Bilan

Ce manuscrit décrit de manière synthétique mon activité de recherche de ces 10 dernières années, depuis mon recrutement au sein l'ISIR en septembre 2008. J'y ai présenté les contributions apportées par les différents étudiants en thèse que j'ai eu la chance de co-encadrer sur deux thématiques :

— **en audition robotique** : présentée au sein du chapitre 2, cette thématique traite de la façon dont un robot, équipé (entre autre) de la capacité de percevoir des sons, peut analyser une scène sonore. Cette problématique a tout d'abord été abordée lors de la thèse de Karim YOUSSEF, dont l'objectif était de proposer des schémas simples d'apprentissage de la localisation d'une source sonore robuste aux changements acoustiques. En parallèle, la thèse d'Alban PORTELLO –qui s'est déroulée au LAAS-CNRS, à Toulouse– a permis de proposer une des premières approches en audition robotique visant à exploiter les conséquences de l'action de la plateforme mobile sur la perception binaurale pour estimer la position de sources sonores dans l'environnement, même en présence de bruits ou lorsqu'elles émettent de manière intermittente. Enfin, la thèse de Benjamin COHEN-LYHVER a été l'occasion de démontrer l'intérêt de la multimodalité pour l'analyse d'une scène, mais également de renforcer le rôle de l'action pour cette tâche via la recherche active de données audio ou visuelles manquantes ;

— **en perception sensorimotrice** : abordée à l'occasion du chapitre 3, cette thématique vise à appréhender le rôle de l'action au sein de la perception, avec l'idée que le flux sensorimoteur dispose en son sein d'éléments –comme une structure ou des invariants– susceptibles de permettre à un agent naïf d'inférer des propriété de son interaction avec son environnement. Ces questions ont d'abord été défrichées à l'occasion de la thèse d'Alban LAFLAQUIÈRE, qui a illustré comment un agent pouvait extraire la dimension de l'espace depuis ses variations sensorielles. Puis ce travail a permis d'intuiter l'existence d'invariants sensorimoteurs susceptible de renseigner l'agent sur son monde extérieure. Sur cette base, le travail de thèse de Valentin MARCEL, toujours en cours, vise à proposer une formalisation mathématique de ces invariants. Les résultats obtenus montrent la force de cette approche formelle et ont d'ores et déjà pu être exploités pour la construction d'une représentation interne du corps d'un agent.

Il est clair que ces deux thématiques sont très différentes, en particulier d'un point de vue méthodologique. Pour autant, j'ai toujours cherché à les aborder sous le prisme du rôle de l'action, qui aura servi de fil rouge tout au long des questions scientifiques que j'ai eu l'occasion de traiter. Il me semble en particulier évident qu'un des enjeux majeur de l'audition robotique réside aujourd'hui en l'intégration de considérations actives. Analyser une scène sonore depuis des capteurs statiques placés au sein d'un environnement connu et maîtrisé

est aujourd'hui en passe d'être un problème résolu [1]. Les approches par apprentissage profond, aidées par des ressources computationnelles importantes et couplées à des quantités de données en augmentation permanente, permettent aujourd'hui d'imaginer interpréter une scène sonore d'une manière beaucoup plus fiable et robuste qu'on ne pouvait le faire il y a seulement 10 ans. Énormément d'applications grand-public bénéficient de ces avancées, et il semble alors naturel de s'imaginer qu'un robot soit doté des mêmes capacités. Mais les données perçues par un robot ne sont pas aussi maîtrisées qu'on le voudrait : bruits, réverbérations, mouvements, etc. tout contribue à complexifier l'analyse. Mais à la différence des approches statiques, pour lesquelles certaines de ces difficultés peuvent être résolues par une exploration maximale de l'espace des données, l'action du robot permet d'envisager une autre approche. Un son est mal reconnu ? Le robot peut s'en approcher, comme le ferait naturellement un humain dans les mêmes conditions. Il existe une ambiguïté sur la localisation d'un événement sonore ? Il suffit de faire bouger légèrement le récepteur acoustique pour la lever, comme nous tournons légèrement la tête pour améliorer notre localisation des sons. L'intérêt de cette approche active est d'autant plus important pour le paradigme binaural qui ne dispose que de deux microphones pour traiter ces données acoustiques difficiles. Enfin, comme illustré dans ce document dans un contexte robotique, utiliser la vision –et de manière plus générale la multimodalité– permet très souvent d'aider à la compréhension que nous avons de notre environnement.

Inclure l'action au sein de l'analyse sonore semble donc être particulièrement pertinent, et le nombre de travaux tâchant d'exploiter ces capacités motrices est en augmentation notable, que ce soit au sein des Communautés Robotique et/ou Signal/Acoustique. Il reste donc sur ce sujet énormément de travail à effectuer. Et ce travail, je souhaite maintenant l'aborder non plus en l'attaquant sous le prisme d'une modalité particulière (audio par exemple, mais j'y reviendrai un peu plus loin), mais dans le cadre plus général qu'est celui de l'approche sensorimotrice de la perception, introduit dans ce manuscrit. C'est à son sujet que je détaille dans la sous section suivante les travaux en cours et perspectives possibles à plus ou moins long terme.

4.2 Travail en cours et perspectives

Emergence d'une représentation interne depuis le flux sensorimoteur : ce travail en cours est traité par V. MARCEL à l'occasion de sa fin de thèse. Il s'agit ici d'aller plus loin dans la formalisation proposée en §3.3.4 et traitant du raffinement de la représentation interne de l'interaction de l'agent avec son environnement. A court terme, il s'agit :

- d'inclure des considérations topologiques nécessitant l'introduction de notions statistiques accessibles à l'agent ; plus précisément, il s'agit de découvrir une structure *empirique* accessible à l'agent, et d'étudier sous quelle(s) hypothèse(s) cette structure conserve les propriétés topologiques *objectives* de l'espace externe à celui-ci ;
- sur la base de ces considérations topologiques attaquées d'une manière très formelle, il faut ensuite proposer une réalisation exploitable algorithmiquement. Ce passage à l'exploitation expérimentale (simulée ou sur des plateformes réelles) nécessite une reformalisation discrète des considérations précédentes. Se pose également des problématiques de robustesse (aux bruits dans les différents capteurs proprio ou extéroceptifs), de répétabilité, etc. pas encore clairement prises en compte dans les différents éléments de formalisation proposés.

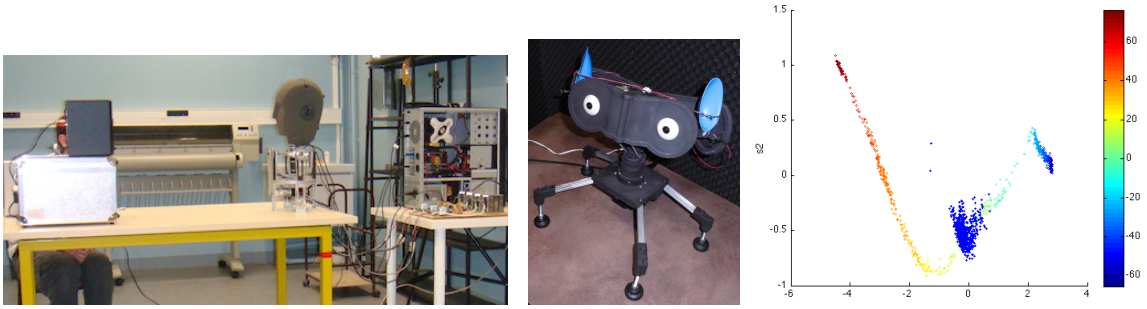
V. MARCEL dispose déjà de résultats concluants sur une partie de ces aspects, et un article est en cours de rédaction pour présenter ces éléments.

1. Que le lecteur inquiet soit rassuré, il reste sans le moindre doute des travaux à mener dans le domaine, en particulier dans les cas multi-sources.

Vers un formalisme incluant les actions pour la prédiction sensorielle : bien que les éléments de formalisation proposés pour rendre compte et interpréter l'information embarquée au sein du flux sensorimoteur soient suffisamment puissants pour permettre, dans des cas bien identifiés, à un agent de construire les contingences de son interaction avec son environnement, nous sommes encore loin d'une théorie sensorimotrice de la perception que se veut complète. Par exemple, tous les agents envisagés sont encore ancrés dans leur environnement : aucun mouvement *global* du robot n'est permis. En particulier, un simple robot à roue pour lequel la proprioception renverrait seulement la position angulaire de chacune de ses roues pose déjà problème : l'information motrice ne donne plus accès à une mesure absolue image d'une position spatiale non ambiguë. A vrai dire, exploiter l'ensemble $\mathcal{M}$ vu comme image des configurations motrices de l'agent pose problème : pour le robot à roue précédent, ou pour un robot humanoïde, à une même configuration motrice dans $\mathcal{M}$ correspond une infinité de positions dans l'espace. Dès lors, les représentations pouvant être obtenues via la formalisation proposée ne peuvent être que locales et relatives ... alors que l'espace et ses propriétés sont a priori identiques en tout point.

Pour résoudre ce problème, je propose de travailler non plus sur l'espace des configurations motrices $\mathcal{M}$, mais plutôt sur un nouvel ensemble $\mathcal{A}$ d'actions *qui agissent sur* $\mathcal{M}$. Par conséquent, une action est maintenant une fonction de $\mathcal{M}$ dans $\mathcal{M}$ et dont on espère que les propriétés conservent les invariants et structures déjà identifiés. Ces actions ainsi formalisées représentent maintenant une *différence* entre deux configurations motrices, différence qui peut être appliquée en n'importe quelle configuration spatiale de l'agent. En ce sens, on peut espérer extraire des informations générales (et non plus locales) de l'interaction de l'agent avec son environnement. C'est exactement ce sur quoi travaille aujourd'hui Jean GODON, qui commence sa thèse sur cette idée. A ce stade, J. GODON a déjà su proposer des éléments de formalisation prometteurs (structure algébrique des actions, découverte de primitives motrices) permettant dans certain cas d'envisager une application concrète de cette théorie pour des tâches de prédiction sensorielle. Ces idées sont à rapprocher des formalismes de codage prédictif (souvent utilisés via des approches bayésiennes) dont on pense qu'ils représentent une part non négligeable des opérations effectuées par notre propre cerveau à des fins d'anticipation. Quelque part, j'y vois également l'occasion de créer un lien inattendu avec les travaux initiaux de B. COHEN-LHYVER en audio sur les notions de saillance, de congruence, et de réaction aux stimuli dont la prédiction est faussée par des entrées sensorielles inattendues.

Et pourquoi pas un peu d'audition sensorimotrice ? Jusqu'à présent, j'ai pris soin de séparer les deux problématiques abordées dans ce manuscrit, pour les raisons évoquées précédemment. Mais ne pourrions nous pas envisager d'exploiter le cadre sensorimoteur pour conduire une analyse audio d'une scène (sonore) ? D'un point de vue conceptuel, rien ne s'y oppose : la formalisation sensorimotrice proposée est a priori suffisamment générique pour s'appliquer à n'importe quelle modalité, dont l'audio. Pour autant, l'audition est fondamentalement différente de la vision dans la mesure où l'audition ne capture pas directement une information spatiale. Sur une caméra, et même si j'ai veillé à ce qu'aucun a priori de ce type ne soit exploité, les pixels disposent d'une organisation spatiale bien identifiée ; en d'autres termes, la vision apparaît comme un sens intrinsèquement spatial au sein duquel des pixels proches codent une infirmation proche spatialement. A l'opposé, il n'y a rien de tel dans le domaine audio : si proximité il y a, c'est dans la répartition fréquentielle de l'information. Et c'est tout l'intérêt que d'aborder cette problématique par le biais de modalités pour lesquelles l'intuition de l'expérimentateur n'est pas aussi simple que pour la vision. Pour en avoir fait moi même l'expérience, cette intuition est parfois trompeuse et souvent réductrice. Travailler dans le domaine audio peut permettre alors d'éviter d'introduire, sans forcément en avoir conscience, un biais ou une information a priori non explicite dans les réflexions.



(A) Estimation de dimension

(B) Robot Binnobot

(C) Représentation basse-dimension

FIGURE 4.1 – Quelques expérimentations évaluant la pertinence du formalisme sensorimoteur dans le domaine audio. (A) Estimation de la dimension de l'espace depuis des données binaurales acquises depuis une tête binaurale KU100. (B) Robot Binnobot, équipé d'une audition binaurale. (C) Représentation basse-dimension de la variété sensorielle audio du robot Binnobot. La variété est de dimension 1, et la couleur correspond à l'angle de la source sonore. Images tirées de (GARCIA et al., 2014) et (LAFLAQUIÈRE, ARGENTIERI, GAS et al., 2010).

Cette envie à terme d'envisager une audition sensorimotrice s'est d'ores et déjà traduite par des contributions ponctuelles. On pourra citer par exemple (LAFLAQUIÈRE, ARGENTIERI, GAS et al., 2010) au sein duquel nous avons exploité la modalité audio depuis une tête binaurale KU100 montée sur un cou capable de tourner autour de trois axes (cf. figure 4.1c) pour estimer la dimension de l'espace selon l'approche présentée en §3.2.2. Ou encore (GARCIA et al., 2014) où nous avons eu l'occasion d'exprimer en termes sensorimoteurs la localisation d'une source sonore dans l'environnement. Sous réserve de disposer d'un réflexe actif visant à faire face à la source à localiser (par minimisation de la différence interaurale en amplitude par exemple), on peut montrer que la variété sensorielle du robot est mono-dimensionnelle et qu'elle capture en son sein la position angulaire de la source sonore, comme représenté figure 4.1c. A la différence d'une approche traditionnelle de localisation de source, qui exprimerait cette localisation en termes objectifs (un azimut relatif à la tête par exemple), la localisation est maintenant exprimée en termes internes moteurs. Le robot apprend ainsi à associer une commande motrice à une sensation audio, sans plus d'a priori. Les performances de cette approche sont aussi bonnes que les approches traditionnelles en environnement contrôlé. Il faudrait néanmoins en évaluer la robustesse dans des contextes bruités et réverbérant, ainsi qu'en situation multisource pour laquelle la multimodalité est absolument nécessaire. On retrouve là l'ensemble des thèmes abordés dans ce manuscrit et qui seront vraisemblablement ceux sur lesquels je me focaliserai, dans un contexte sensorimoteur, les prochaines années.

Bibliographie

AFGHAH, T., A. ALLEN, P. OTTO et A. J. BENJAMIN (2017). « The Evaluation of the Effect of Sound Directionality in Horizontal Plane on the Human Auditory Distance Perception in a Large Reverberant Room ». In : *Audio Engineering Society Convention 142*.

AHISSAR, M. et S. HOCHSTEIN (2004). « The Reverse Hierarchy Theory of Visual Perceptual Learning ». In : *Trends in Cognitive Sciences* 8.10, p. 457–64. ISSN : 1364-6613.

ALLEN, J. B. et D. A. BERKLEY (1979). « Image method for efficiently simulating small-room acoustics ». In : *Journal of the Acoustical Society of America* 65.4.

ALOIMONOS, Y., I. WEISS et A. BANDOPADHAY (1988). « Active vision ». In : *Int. Journ. of Computation Vision* 1.4, p. 333–356.

ARGENTIERI, S. et P. DANÈS (2007). « Broadband variations of the MUSIC high-resolution method for Sound Source Localization in Robotics ». In : *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, p. 2009–2014. ISBN : 2153-0858.

ARGENTIERI, S., P. DANES et P. SOUÈRES (2006). « Modal Analysis Based Beamforming for Nearfield or Farfield Speaker Localization in Robotics ». In : *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, p. 866–871. ISBN : 2153-0858.

ARGENTIERI, S., P. DANÈS et P. SOUÈRES (2015). « A survey on sound source localization in robotics : From binaural to array processing methods ». In : *Computer Speech & Language* 34.1, p. 87–112. ISSN : 0885-2308.

ARGENTIERI, S., A. PORTELLO, M. BERNARD, P. DANÈS et B. GAS (2013). « Binaural Systems in Robotics ». In : *The Technology of Binaural Listening*. Sous la dir. de J. BLAUERT. Modern Acoustics and Signal Processing. Berlin, Heidelberg : Springer. Chap. 9, p. 225–253.

BACH-Y-RITA, P., C. C. COLLINS, F. SAUNDERS, B. WHITE et L. SCADDEN (1969). « Vision Substitution by Tactile Image Projection ». In : 221, p. 963–4.

BACH-Y-RITA, P. et S. W. KERCEL (2003). « Sensory substitution and the human–machine interface ». In : *Trends in Cognitive Sciences* 7.12, p. 541–546.

BAJCSY, R. (1988). « Active perception ». In : *Proc. of IEEE* 76.8, p. 996–1005.

BEDNAR, A., F. M. BOLAND et E. C. LALOR (2017). « Different spatio-temporal electroencephalography features drive the successful decoding of binaural and monaural cues for sound localization ». In : *European Journal of Neuroscience* 45.5, p. 679–689. ISSN : 1460-9568.

BENESTY, J., J. CHEN et Y. HUANG (2008). *Microphone Array Signal Processing*. Springer Topics in Signal Processing. Springer.

BERGMAN, A. (1990). *Auditory Scene Analysis : The Perceptual Organization of Sound*. MIT Press.

BERNARD, M., P. PIRIM, A. de CHEVEIGNÉ et B. GAS (2012). « Sensorimotor learning of sound localization from an auditory evoked behavior ». In : *Robotics and Automation (ICRA), IEEE International Conference on*, p. 91–96.

BERNARD, M. (2014). « Audition active et intégration sensorimotrice pour un robot autonome bioinspiré ». Thèse de doct. Université Pierre et Marie Curie.

BERNSTEIN, L. R. et C. TRAHIOTIS (1996). « The normalized correlation : Accounting for binaural detection across center frequency ». In : *Journal of the Acoustical Society of America* 100.6.

BESSIERE, P., C. LAUGIER et R. SIEGWART (2008). *Probabilistic Reasoning and Decision Making in Sensory-Motor Systems*. Springer.

BOHG, J., K. HAUSMAN, B. SANKARAN, O. BROCK, D. KRAGIC, S. SCHAAL et G. S. SUKHATME (2017). « Interactive Perception : Leveraging Action in Perception and Perception in Action ». In : *IEEE Transactions on Robotics* 33.6, p. 1273–1291.

BONDY, J.-A. et U. S. R. MURTY (2007). *Graph theory*. Graduate texts in mathematics. New York, London : Springer. ISBN : 978-1-8462-8969-9.

BONNAL, J., S. ARGENTIERI, P. DANÈS, J. MANHÈS, P. SOUÈRES et M. RENAUD (2010). « The EAR Project ». In : *Journal of the Robotics Society of Japan (RSJ), Special issue "Robot Audition", Invited paper* 28.1, p. 10–13.

BRETTE, R. (2013). « Subjective Physics ». In : *arXiv :1311.3129*.

BRONKHORST, A. (2000). « The Cocktail Party Phenomenon : A Review of Research on Speech Intelligibility in Multiple-Talker Conditions ». In : *Acta Acustica united with Acustica* 86, p. 117–128.

BUSTAMANTE, G., P. DANÉS, T. FORGUE et A. PODLUBNE (2016). « Towards information-based feedback control for binaural active localization ». In : *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p. 6325–6329.

BUSTAMANTE, G., P. DANÈS, T. FORGUE, A. PODLUBNE et J. MANHÈS (2017). « An information based feedback control for audio-motor binaural localization ». In : *Autonomous Robots*. ISSN : 1573-7527.

CAMPBELL, D. R., K. PALOMÄKI et G. BROWN (2005). « A MATLAB simulation of "shoebox" room acoustics for use in research and teaching ». In : *Computer Information Systems* 9.3.

CARTER, G. C., A. H. NUTTALL et P. G. CABLE (1973). « The smoothed coherence transform ». In : *Proceedings of the IEEE* 61.10, p. 1497–1498. ISSN : 0018-9219. DOI : 10.1109/PROC.1973.9300.

CHAUMETTE, F. (1998). « De la perception à l'action : l'asservissement visuel, de l'action à la perception : la vision active. » Habilitation à diriger des recherches. Université de Rennes 1.

COHEN-LHYVER, B., S. ARGENTIERI et B. GAS (2015). « Modulating the auditory turn-to reflex on the basis of multimodal feedback loops : The Dynamic Weighting model ». In : *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO), p. 1109–1114.

COHEN-LHYVER, B. (2017). « Modulation de mouvements de tête pour l'analyse multimodale d'un environnement inconnu ». Thèse de doct. Université Pierre et Marie Curie.

COHEN-LHYVER, B., S. ARGENTIERI et B. GAS (2016). « Multimodal fusion and inference using binaural audition and vision ». In : *International Conference on Acoustics*, ICA2016–0715.

COHEN-LYVER, B., S. ARGENTIER et B. GAS (2018). « The Head Turning Modulation system : an active multimodal paradigm for intrinsically motivated exploration of unknown environments ». In : *Frontiers in Neurobotics*.

DAPOGNY, A., K. BAILLY et S. DUBUISSON (2018). « Dynamic Pose-Robust Facial Expression Recognition by Multi-View Pairwise Conditional Random Forests ». In : *IEEE Transactions on Affective Computing*. to appear.

DELEFORGE, A. et R. HORAUD (2012). « The Cocktail Party Robot : Sound Source Separation and Localisation with an Active Binaural Head ». In : HRI 2012 - 7th ACM/IEEE International Conference on Human Robot Interaction. Boston, United States, p. 431–438.

DEMARTINES, P. et J. HERAULT (1997). « Curvilinear component analysis : a self-organizing neural network for nonlinear mapping of data sets ». In : *IEEE Transactions on Neural Networks* 8.1, p. 148–154.

DIJKSTRA, E. W. (1959). « A note on two problems in connexion with graphs ». In : *Numerische Mathematik* 1.1, p. 269–271. ISSN : 0945-3245. DOI : 10.1007/BF01386390.

DRIVER, J. et C. SPENCE (1998). « Attention and Cross Modal Construction of Space ». In : *Trends in Cognitive Sciences* 2.7, p. 254–262.

DUANGUDOM, V. et D. V. ANDERSON (2007). « Using Auditory Saliency to Understand Complex Auditory Scenes ». In : *European Signal Processing Conference*. 15th.

EVERS, C., Y. DORFAN, S. GANNOT et P. A. NAYLOR (2017). « Source tracking using moving microphone arrays for robot audition ». In : *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p. 6145–6149.

FALLER, C. et J. MERIMAA (2004). « Source Localization in Complex Listening Situations : Selection of Binaural Cues based on Interaural Coherence ». In : *Journal of the Acoustical Society of America* 116.5.

GAPENNE, O. (2014). « The co-constitution of the self and the world : action and proprioceptive coupling ». In : *Frontiers in Psychology* 5.

GARCIA, B., M. BERNARD, S. ARGENTIERI et B. GAS (2014). « Sensorimotor Learning of Sound Localization for an Autonomous Robot ». In : *Forum Acusticum*, PACS no. 43.60.Np, 43.66.Pn.

GAS, B. et S. ARGENTIERI (2016). « Une brève introduction à la perception sensori-motrice en robotique ». In : *Intellectica* 65.2016/1, p. 27–61.

GIBSON, J. J. (1979). *The ecological approach to visual perception*. Boston : Houghton.

GRONDIN, F., D. LÉTOURNEAU, F. FERLAND, V. ROUSSEAU et F. MICHAUD (2013). « The ManyEars open framework ». In : *Autonomous Robots* 34.3, p. 217–232. ISSN : 1573-7527. DOI : 10.1007/s10514-012-9316-x. URL : https://doi.org/10.1007/s10514-012-9316-x.

H. D. KIM, K. KOMATANI, T. OGATA et H. G. OKUNO (2008). « Design and evaluation of two-channel-based sound source localization over entire azimuth range for moving talkers ». In : *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, p. 2197–2203. ISBN : 2153-0858.

HANNAN, E. J. et P. J. THOMSON (1973). « Estimating group delay ». In : *Biometrika* 60.2, p. 241–253. DOI : 10.1093/biomet/60.2.241.

HELD, R. et A. HEIN (1963). « Movement-produced stimulation in the development of visually guided behavior ». In : *J Comp Physiol Psychol*, p. 872–876.

HEUDIN, J.-C. (2008). *Les Créatures artificielles : des automates aux mondes virtuels*. Odile Jacob.

HIOKA, Y., K. NIWA, S. SAKAUCHI, K. FURUYA et Y. HANEDA (2011). « Estimating Direct-to-Reverberant Energy Ratio Using D/R Spatial Correlation Matrix Model ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 19.8, p. 2374–2384. ISSN : 1558-7916. DOI : 10.1109/TASL.2011.2134091.

HUTTENLOCHER, D. P., G. A. KLANDERMAN et W. J. RUCKLIDGE (1993). « Comparing images using the Hausdorff distance ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15.9, p. 850–863.

INCE, G. (2011). « Ego Noise Estimation for Robot Audition ». Thèse de doct. Tokyo Institute of Technology.

K. YOUSSEF, S. ARGENTIERI et J. L. ZARADER (2012). « Towards a systematic study of binaural cues ». In : *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, p. 1004–1009. ISBN : 2153-0858.

— (2013). « A learning-based approach to robust binaural sound localization ». In : *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, p. 2927–2932. ISBN : 2153-0858.

KALINLI, O. et S. NARAYANAN (2007). « A Saliency-Based Auditory Attention Model with Applications to Unsupervised Prominent Syllable Detection in Speech ». In : *Interspeech*, p. 1–4.

KATZ, B. F. G. et M. NOISTERNIG (2014). « A comparative study of interaural time delay estimation methods ». In : *The Journal of the Acoustical Society of America* 135.6, p. 3530–3540. ISSN : 0001-4966. (Visité le 20/11/2017).

KAYSER, C., C. I. PETKOV, M. LIPPERT et N. K. LOGOTHETIS (2005). « Mechanisms for Allocating Auditory Attention : An Auditory Saliency Map ». In : *Current Biology* 15, p. 1943–1947.

KELLERMANN, W. (2016). *Embodied Audition for RobotS (EARS) project*. Final report. URL : https://robot-ears.eu/.

KHAMASSI, M. et S. DONCIEUX (2016). « Nouvelles approches en Robotique Cognitive ». In : *Intellectica* 65.2016/1, p. 7–25.

KINSLER, L. E., A. R. FREY, A. B. COPPENS et J. V. SANDERS (1999). *Fundamentals of Acoustics, 4th Edition*, p. 560.

KNEIP, L. et C. BAUMANN (2008). « Binaural model for artificial spatial sound localization based on interaural time delays and movements of the interaural axis ». In : *The Journal of the Acoustical Society of America* 124.5, p. 3108–3119.

KNILL, D. et W. RICHARDS (1996). *Perception as Bayesian Inference*. Cambridge University Press.

KOHLRAUSCH, A., J. BRAASCH, D. KOLOSSA et J. BLAUERT (2013). « An introduction to binaural processing ». In : *The Technology of Binaural Listening*. Springer, Berlin, Heidelberg. Modern Acoustics and Signal Processing. Blauert, Jens, p. 1–32.

KOLOSSA, D. et G. BROWN (2016). *The TWO !EARS project. Final report and evaluated software for analysis of dynamic auditory scenes*. Deliverable. URL : https://twoears.eu/.

KUMON, M., K. FUKUSHIMA, S. KUNIMATSU et M. ISHITOBI (2010). « Motion planning based on simultaneous perturbation stochastic approximation for mobile auditory robots ». In : *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, p. 431–436. DOI : 10.1109/IROS.2010.5649244.

LAFLAQUIÈRE, A., S. ARGENTIERI, O. BREYSSE, S. GENET et B. GAS (2012). « A non-linear approach to space dimension perception by a naive agent ». In : *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, p. 3253–3259.

LAFLAQUIÈRE, A., S. ARGENTIERI, B. GAS et E. CASTILLO-CASTENADA (2010). « Space dimension perception from the multimodal sensorimotor flow of a naive robotic agent ». In : *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, p. 1520–1525. DOI : 10.1109/IROS.2010.5651695.

LAFLAQUIÈRE, A. (2013). « Approche sensorimotrice de la perception de l'espace pour la robotique autonome ». Thèse de doct. Université Pierre et Marie Curie.

LAFLAQUIÈRE, A., J. K. O'REGAN, S. ARGENTIERI, B. GAS et A. V. TEREKHOV (2015). « Learning agent's spatial configuration from sensorimotor invariants ». In : *Robotics and Autonomous Systems* 71, p. 49–59.

LAFLAQUIÈRE, A., J. O'REGAN, B. GAS et A. TEREKHOV (2018). « Discovering space - Grounding spatial topology and metric regularity in a naive agent's sensorimotor experience ». In : *Neural Networks*, in press. ISSN : 0893-6080.

LE MEUR, O., P. LE CALLET, D. BARBA et D. THOREAU (2006). « A Coherent Computational Approach to Model Bottom-Up Visual Attention ». In : *Transactions on Pattern Analysis and Machine Intelligence* 28, p. 802–817.

LEE, J. A. et M. VERLEYSEN (2007). *Nonlinear Dimensionality Reduction*. 1st. Springer Publishing Company, Incorporated.

LIU, R. et Y. WANG (2010). *Azimuthal source localization using interaural coherence in a robotic dog : Modeling and application*. T. 28. DOI : 10.1017/S0263574709990865.

LÖLLMANN, H. W., A. MOORE, P. A. NAYLOR, B. RAFAELY, R. HORAUD, A. MAZEL et W. KELLERMANN (2017). « Microphone array signal processing for robot audition ». In : *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, p. 51–55.

LU, Y. C. et M. COOKE (2010). « Binaural Estimation of Sound Source Distance via the Direct-to-Reverberant Energy Ratio for Static and Moving Sources ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 18.7, p. 1793–1805. ISSN : 1558-7916. DOI : 10.1109/TASL.2010.2050687.

LU, Y.-C. et M. COOKE (2010). « Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners ». In : *Speech Communication*.

LUNATI, V., J. MANHÈS et P. DANÈS (2012). « A versatile System-on-a-Programmable-Chip for array processing and binaural robot audition ». In : *IEEE/RSJ International Conference on Intelligent Robots and Systems*, p. 998–1003.

MAGASSOUBA, A., N. BERTIN et F. CHAUMETTE (2015). « Sound-based control with two microphones ». In : *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, p. 5568–5573.

MAGASSOUBA, A. (2016). « Aural servo : towards an alternative approach to sound localization for robot motion control ». Thèse de doct.

MAILLARD, M., O. GAPENNE, L. HAFEMEISTER et P. GAUSSIER (2005). « Perception as a dynamical sensori-motor attraction basin ». In : *Adances in Artificial Life, ECAL 2005*. T. LNAI 3630, p. 37–46.

MARCEL, V., S. ARGENTIERI et B. GAS (2017). « Building a Sensorimotor Representation of a Naive Agent's Tactile Space ». In : *IEEE Transactions on Cognitive and Developmental Systems* 9.2, p. 141–152.

MARKOVIC, I., A. PORTELLO, P. DANES, I. PETROVIC et S. ARGENTIERI (2013). « Active speaker localization with circular likelihoods and bootstrap filtering ». In : *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, p. 2914–2920. DOI : 10.1109/IROS.2013.6696769.

MARR, D. (1982). *Vision : A Computational Investigation into the Human Representation and Processing of Visual Information*. New York : Freeman.

MARTINSON, E., T. APKER et M. BUGAJSKA (2011). « Optimizing a reconfigurable robotic microphone array ». In : *Intelligent Robots and Systems (IROS), IEEE/RSJ International Conference on*, p. 125–130.

MEYER, J.-A. (2015). *Dei ex Machinis : Volume I – De l'Antiquité à Hans Schlottheim*. Les édition du Net.

MIDDLEBROOKS, J. et D. M. GREEN (1991). *Sound Localization by Human Listeners*. T. 42.

MOLHOLM, S., A. MARTINEZ, W. RITTER, D. C. JAVITT et J. J. FOXE (2005). « The neural circuitry of pre-attentive auditory change-detection : An fMRI study of pitch and duration mismatch negativity generators ». In : *Cerebral Cortex* 15.5, p. 545–551. ISSN : 10473211.

NÄÄTÄNEN, R., A. GAILLARD et S. MÄNTYSALO (1978). « Early Selective-Attention Effect on Evoked Potential Reinterpreted ». In : *Acta Psychologica* 42, p. 313–329.

NAKADAI, K., H. G. OKUNO, H. NAKAJIMA, Y. HASEGAWA et H. TSUJINO (2008). « An open source software system for robot audition HARK and its evaluation ». In : *Humanoids 2008 - 8th IEEE-RAS International Conference on Humanoid Robots*, p. 561–566.

NAKADAI, K., T. LOURENS, H. G. OKUNO et H. KITANO (2000). « Active Audition for Humanoid ». In : *National Conference on Artificial Intelligence*. Austin, Texas, p. 832–839.

NELKEN, I. et M. AHISSAR (2006). « High-Level and Low-Level Processing in the Auditory System : The Role of Primary Auditory Cortex ». In : *Dynamic of Speech Production and Perception*, p. 5–12.

NOTHDURFT, H.-C. (2006). « Salience and Target Selection in Visual Search ». In : *Visual Cognition* 14.4-8, p. 514–542.

ODO, W., D. KIMOTO, M. KUMON et T. FURUKAWA (2017). *Active sound source localization by Pinnae with recursive Bayesian estimation*. T. 29.

OKUTANI, K., T. YOSHIDA, K. NAKAMURA et K. NAKADAI (2012). « Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter ». In : *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, p. 3288–3293.

OLIVA, A., A. TORRALBA, M. S. CASTELHANO et J. M. HENDERSON (2003). « Top-Down Control of Visual Attention in Object Detection ». In : *IEEE International Conference on Image Processing, September 14-17* 1, p. 1–4. ISSN : 1522-4880.

O'REGAN, J. K. et A. NOË (2001). « A sensorimotor account of vision and visual consciousness ». In : *Behavioral and Brain Sciences* 24.5, p. 939–1031.

O'REGAN, J. (2011). *Why Red Doesn't Sound Like a Bell : Understanding the feel of consciousness*. OUP USA. ISBN : 9780199775224.

PETERS II, R. A., K. E. HAMBUCHEN, K. KAWAMURA et D. M. WILKES (2001). « The Sensory Ego-Sphere as a Short-Term Memory for Humanoids ». In : *Proceedings of the IEEE-RAS International Conference on Humanoid Robots* 1, p. 451–459.

PHILIPONA, D. (2008). « Développement d'un cadre mathématique pour la notion de dépendances sensorimotrices, dans le contexte d'une théorie de l'expérience sensorielle ». Thèse de doct. Ecole Normale Supérieure de Cachan.

PHILIPONA, D., J. K. O'REGAN et J.-P. NADAL (2003). « Is there something out there ? : Inferring space from sensorimotor dependencies ». In : *Neural Computation* 15.9, p. 2029–2049. ISSN : 0899-7667. DOI : 10.1162/089976603322297278.

PHILIPONA, D. et J. KEVIN O'REGAN (2006). « Color naming, unique hues, and hue cancellation predicted from singularities in reflection properties ». In : 23, p. 331–9.

POINCARÉ, H. (1887). « Sur les hypothèses fondamentales de la géométrie ». In : *Bulletin de la Société Mathématique de France* 15, p. 203–216.

— (1895). « On the Foundations of Geometry ». In : *The Monist* 9, p. 1–43.

— (1902). *La science et l'hypothèse*. Flammarion.

— (1913). *Dernières pensées*. Flammarion.

PORTELLO, A., P. DANÈS et S. ARGENTIERI (2011). « Acoustic models and Kalman filtering strategies for active binaural sound localization ». In : *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, p. 137–142. DOI : 10.1109/IROS.2011.6094842.

PORTELLO, A., P. DANÈS, S. ARGENTIERI et S. PLEDEL (2013). « HRTF-based source azimuth estimation and activity detection from a binaural sensor ». In : *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, p. 2908–2913. DOI : 10.1109/IROS.2013.6696768.

PORTELLO, A. (2013). « Localisation Binaurale Active de Sources Sonores en Robotique Humanoïde ». Thèse de doct. Université Paul Sabatier, Toulouse 3.

PORTELLO, A., G. BUSTAMANTE, P. DANÈS, J. PIAT et J. MANHÈS (2014). « Active localization of an intermittent sound source from a moving binaural sensor ». In : *Forum Acusticum*. T. PACS no. 43.60.Jn, 43.66.Pn. Krakow.

RAAKE, A. (2016). *The TWO !EARS project*. Final report. URL : https://twoears.eu/.

RAYLEIGH, L. (1907). « XII. On our perception of sound direction ». In : *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 13.74, p. 214–232. ISSN : 1941-5982.

RICHTERS, D. P. et R. T. ESKEW Jr. (2009). « Quantifying the effect of natural and arbitrary sensorimotor contingencies on chromatic judgments ». In : *Journal of Vision* 9.4, p. 27.

ROSCHIN, V. Y. et A. FROLOV (2011). « A Neural Network Model for the Acquisition of a Spatial Body Scheme Through Sensorimotor Interaction ». In : *Neural Computation* 23, p. 1821–1834.

ROTH, P. R. (1971). « Effective measurements using digital signal analysis ». In : *IEEE Spectrum* 8.4, p. 62–70. ISSN : 0018-9235.

RUESCH, J., M. LOPES, A. BERNARDINO, J. HÖRNSTEIN, J. SANTOS-VICTOR et R. PFEIFER (2008). « Multimodal Saliency-Based Bottom-Up Attention a Framework for the Humanoid Robot iCub ». In : *Proceedings - IEEE International Conference on Robotics and Automation*, p. 962–967. ISSN : 10504729.

RUSSELL, S. J. et P. NORVIG (2003). *Artificial Intelligence : A Modern Approach*. 2e éd. Pearson Education. ISBN : 0137903952.

SHAMMA, S. (2008). « On the Emergence and Awareness of Auditory Objects ». In : *PLoS biology* 6.6, e155. ISSN : 1545-7885.

SU, D., K. NAKAMURA, N. K. et J. V. MIRO (2016). « Robust sound source mapping using three-layered selective audio rays for mobile robots ». In : *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, p. 2771–2777.

T. MAY, N. MA et G. J. BROWN (2015). « Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues ». In : *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 2679–2683. ISBN : 1520-6149.

T. MAY, S. VAN DE PAR et A. KOHLRAUSCH (2011). « A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 19.1, p. 1–13. ISSN : 1558-7916.

— (2012). « A Binaural Scene Analyzer for Joint Localization and Recognition of Speakers in the Presence of Interfering Noise Sources and Reverberation ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 20.7, p. 2016–2030. ISSN : 1558-7916.

TENENBAUM, J. B., V. d. SILVA et J. C. LANGFORD (2000). « A Global Geometric Framework for Nonlinear Dimensionality Reduction ». In : *Science* 290.5500, p. 2319–2323.

THEILER, J. (1990). « Estimating fractal dimension ». In : *J. Opt. Soc. Am. A* 7.6, p. 1055–1073.

TREISMAN, A. M. et G. GELADE (1980). « A Feature-Integration Theory of Attention ». In : *Cognitive psychology* 12.1, p. 97–136.

V. NGUYEN, Q., F. COLAS, E. VINCENT et F. CHARPILLET (2017). « Long-Term Robot Motion Planning for Active Sound Source Localization with Monte Carlo Tree Search ». In : *Hands-free Speech Communication and Microphone Arrays*. San Francisco, United States.

VAN TREES, H. L. (2002). *Optimum Array Processing*. T. IV. Detection, Estimation, and Modulation Theory. John Wiley & Sons, Inc.

VESA, S. (2009). « Binaural Sound Source Distance Learning in Rooms ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 17.8, p. 1498–1507. ISSN : 1558-7916. DOI : 10.1109/TASL.2009.2022001.

VINCENT, E., A. SINI et F. CHARPILLET (2015). « Audio source localization by optimal control of a mobile robot ». In : *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5630–5634. ISBN : 1520-6149.

WANG, F., Y. TAKEUCHI, N. OHNISHI et N. SUGIE (1997). « A Mobile Robot With Active Localization and Discrimination of a Sound Source ». In : *Journal of the Robotics Society of Japan* 15.2, p. 229–229.

YOUSSEF, K., S. ARGENTIERI et J. L. ZARADER (2012). « A binaural sound source localization method using auditive cues and vision ». In : *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 217–220. DOI : 10.1109/ICASSP.2012.6287856.

YOUSSEF, K. (2013). « Perception binaurale pour l'analyse de scène auditive en robotique ». Thèse de doct. Université Pierre et Marie Curie. (Visité le 02/10/2017).

ZAHORIK, P. (2002). « Assessing auditory distance perception using virtual acoustics ». In : *The Journal of the Acoustical Society of America* 111.4, p. 1832–1846. DOI : 10.1121/1.1458027.

ZHONG, X., L. SUN et W. YOST (2016). « Active binaural localization of multiple sound sources ». In : *Robotics and Autonomous Systems* 85 (Supplement C), p. 83–92.

Résumé

Habilitation à Diriger des Recherches
Un (petit) pas vers la perception interactive

Sylvain ARGENTIERI

La capacité à percevoir et analyser une scène sonore a été identifiée il y a maintenant 20 ans parmi les sept défis majeurs de l'intelligence artificielle en Robotique, et est aujourd'hui une problématique à part entière abordée par de multiples Communautés scientifiques (Traitement de Signal, Acoustique, Robotique). Si la Communauté de l'audition en Robotique s'est particulièrement focalisée sur l'emploi de méthodes dites "d'antennerie", l'utilisation de capteurs binauraux –reproduisant les méthodes de captation de l'information sonore chez l'homme– reste aujourd'hui un défit. Les travaux présentés dans ce manuscrit traitent de ce dernier paradigme, avec une emphase particulière sur la problématique de localisation de source et d'analyse *active* de scènes multimodales dans un contexte réaliste en robotique, intégrant des environnements possiblement bruyants et réverbérants. L'action de la plateforme robotique joue dans ces travaux un rôle majeur : le mouvement du robot, s'il conduit malheureusement à des changements de situation acoustique préjudiciables, peut être également exploité en fermant la traditionnelle boucle action/perception afin d'aider à l'analyse de la scène. L'ensemble de nos contributions sur ce thème de l'audition active en robotique est synthétisé dans une première partie.

En parallèle de ces travaux traitant de la modalité audio dans un cadre actif, le rôle plus formel de l'action dans le processus de perception est également abordé dans ce manuscrit. Il s'agit ici de remettre l'action au centre de l'acte perceptif, au sens où le système ne peut plus strictement percevoir sans agir sur/dans son environnement : on parle alors de *perception interactive*. Il s'agit donc de comprendre comment un système sans aucune connaissance a priori est capable de se construire une représentation de son interaction avec son environnement via la découverte d'invariants au sein de son flux sensorimoteur. Ceci s'opère via l'analyse des conséquences sensorielles d'un mouvement moteur et la découverte progressive de la notion d'espace, partagée à la fois par le système et les objets présents dans son environnement. Il s'agit ici clairement d'un axe de recherche fondamentale, pas nécessairement dédié aux aspects audio (binauraux), avec l'objectif à long terme de proposer une nouvelle façon d'envisager l'action comme support de la perception.

Being able to perceive and analyse an auditory scene has been identified about 20 years ago as one of the seven challenges faced by artificial intelligence in Robotics. It is today a scientific topic on its own, dealt with by multiple scientific Communities (Signal Processing, Acoustics, Robotics). While the Robot Audition community has been mainly focused on microphone array based approaches, exploiting only two microphones in a binaural setup is still a challenging task. The work presented in the first part of this dissertation deals with binaural audition, and is dedicated to sound source localization and *active* multimodal scene analysis in realistic robotics conditions involving noise and reverberations. The movement of the robotic plateform plays a fundamental role in these works : while causing changes in acoustic conditions, the robot action can also been exploited by closing the traditional perception/action loop to better the multimodal scene analysis.

The second part of this dessertation is dedicated to a more formal approach to perception, where action can not be separated from perception anymore : perception is only possible by interacting on and with the environment. This *interactive perception* paradigm allows to study how a naive system is able to build by itself a representation of its interaction with its environmenent by discovering invariant structures inside its own sensorimotor flow. This formal approach could allow the robot to incrementally experience the notion of space, shared by the system and objects in the environment. Such a fundamental problematic has not been addressed specifically inside the audio modality, and aims at proposing a new sensorimotor framework to better understand the perception process that could allow Robotics system to gain in Autonomy.