Audition active et intégration sensorimotrice pour un robot autonome bioinspiré

Mathieu Bernard

Thèse

Présentée à

L'Université Pierre et Marie Curie

École Doctorale SMAER

Par Mathieu BERNARD

Pour obtenir le grade de docteur

Spécialité Robotique

Audition active et intégration sensorimotrice pour un robot autonome bioinspiré

---

soutenue le 15 mai 2014

devant le jury composé de :

M. Philippe Gaussier	Professeur, ETIS/Université de Cergy-Pontoise	Rapporteur
M. Kevin O'Regan	Directeur de Recherche, LPP/CNRS	Rapporteur
M. Angelo Arleo	Directeur de Recherche, NPA/UPMC	Examinateur
M. Radu Horaud	Directeur de Recherche, INRIA	Examinateur
M. Bruno Gas	Professeur, ISIR/UPMC	Directeur
M. Alain de Cheveigné	Directeur de Recherche, LPP/CNRS	Directeur
M. Patrick Pirim	Brain Vision Systems	Invité

Brain Vision Systems
23 rue du Dessous des Berges
75013 Paris

École doctorale SMAER
Université Pierre et Marie Curie
CC 270 - 4 place Jussieu
75252 Paris cedex 05

Institut des Systèmes Intelligents et de Robotique
Université Pierre et Marie Curie - CNRS UMR 7222
CC 173 - 4 place Jussieu
75252 Paris cedex 05

Laboratoire de Psychologie de la Perception
Université Paris Descartes - CNRS UMR 8158
29 rue d'Ulm
75005 Paris

8 avril 2014

En mémoire de Dominique,
en mémoire d'Anne-Lise.

Active audition and sensorimotor integration for a bioinspired autonomous robot

Abstract

The vast majority of perceptual systems proposed in robotics inherit a passive conception of perception, in which the generation of a motor command is the final stage of successive passive processes. In the field of sound sources localization, which is a fundamental task of the auditory system, this passive approach provides good results when the environment is well known and easily modeled. However, difficulties arise when the environment becomes more complex, unknown or changing. These difficulties are a major issue in the field of machine hearing.

This thesis considers a radically different approach inspired by the psychology of perception and theory of sensorimotor contingencies. This approach places action at the heart of the process of perception, which is seen as an interaction of a biological or robotic agent with it's environment. While passive approach requires environmental knowledge, implicitly integrated into models by the robotisist, the sensorimotor approach suggests that this knowledge is acquired by the agent by itself, through its sensorimotor experience.

Thus, this thesis applies the theory of sensorimotor contingencies to sound sources localization for autonomous robots. Based on a model of the auditory system adapted to robotics, this thesis proposes a redefinition of the localization problem in sensorimotor terms. A sensorimotor model of localization is then proposed. It is based on active and low-level perception skills which are used to learn a representation of the auditory space. This representation is then used for a passive localization of new sound sources. By exploiting the active capabilities of the robot, this model eliminates the environment dependencies that put difficulty in the passive approach, thus offering a degree of autonomy higher than current models.

Keywords

Active perception, bioinspired, binaural, sound sources localization, autonous learning, sensorimotor integration.

Résumé

La grande majorité des systèmes perceptifs proposés en robotique héritent d'une conception passive de la perception dans laquelle la génération d'une commande motrice est l'étape ultime d'une succession de traitements purement passifs. Dans le cadre de la localisation de sources sonores, qui est une tâche fondamentale du système auditif, cette approche passive offre de bons résultats lorsque les conditions environnementales sont bien connues et facilement modélisables. Cependant des difficultés apparaissent lorsque l'environnement se complexifie, *a fortiori* s'il est inconnu ou changeant. Ces difficultés constituent un enjeu important dans le domaine de l'audition artificielle.

Cette thèse considère une approche radicalement différente de l'approche passive, inspirée de la psychologie de la perception et de la théorie des contingences sensorimotrices. Cette approche place l'action au coeur du processus de perception, qui est alors vu comme une interaction qu'un agent biologique ou robotique entretient avec son environnement. Alors que l'approche passive nécessite des connaissances sur l'environnement, implicitement intégrées dans les traitements par le roboticien, l'approche sensorimotrice suggère au contraire que ces connaissances sont acquises par l'agent de manière autonome, à travers son expérience sensorimotrice.

Ainsi cette thèse applique la théorie des contingences sensorimotrices à la localisation de sources sonores pour la robotique autonome. Sur la base d'un modèle bioinspiré du système auditif adapté au contexte robotique, cette thèse propose une redéfinition du problème de la localisation en termes sensorimoteurs. Un modèle de localisation sensorimotrice est alors proposé. Celui-ci se base sur des capacités de perception active bas-niveau pour construire une représentation de l'espace auditif qui est ensuite utilisée pour une localisation passive. En exploitant les capacités d'action du robot, ce modèle permet de s'affranchir des dépendances à l'environnement qui mettent en difficulté l'approche passive, en proposant ainsi un degré d'autonomie supérieur à celui des modèles actuels.

Mots clefs

Perception active, bioinspiré, binaural, localisation de sources sonores, apprentissage autonome, intégration sensorimotrice.

Table des matières

Table des figures

Table des abbréviations

k-ppv	k plus proches voisins
CAN	Convertisseur analogique - numérique
CCA	*Curvilinear componant analysis* - analyse en composantes curvilignes
CCE	cellule ciliée externe
CCI	cellule ciliée interne
CN	*Cochlear nucleus* - noyau cochléaire
CNA	Convertisseur numérique - analogique
DCN	*Dorsal cochlear nucleus* - noyau cochléaire dorsal
DOF	*Degree of freedom* - degré de liberté
DSP	*Digital signal processor* - processeur de signal numérique
EM	Espérance-maximisation
ERB	*equivalent rectangular bandwidth* - bande passante rectangulaire équivalente
FFT	*Fast Fourier transform* - transformée de Fourier rapide
GCC	*Generalized cross-correlation* - corrélation croisée généralisée
GCC-PHAT	*GCC with phase transform* - GCC avec transformée de phase
GMM	*Gaussian mixture model* - modèle de mélanges gaussiens
HRTF	*Head related transfer function* - fonction de transfert relative à la tête
HWR	*Half wave rectification* - rectification de demi-onde
IC	*Inferior colliculus* - colliculus inférieur
ICC	*IC central nucleus* - noyau central du IC
ICX	*IC extern nucleus* - noyau externe du IC
ILD	*Interaural level difference* - différence interaurale d'intensité
IPD	*Interaural phase difference* - différence interaural de phase
ITD	*Interaural time difference* - différence interaural de temps d'arrivée
LE	*Laplacian eigenmaps* - cartes propres Laplaciennes
LSO	*Lateral superior olive* - olive supérieure latérale
LTSA	*Local tangent space alignement* - alignement des espaces tangents locaux

MEMS *Microelectromechanical systems* - microsystème électromécanique

MGB *Medial geniculate body* - corps genouillé médial

MLP *Multilayer perceptron* - perceptron multicouches

MODD *Moddemeijer delay criterion* - critère de délai de Moddemeijer

MSO *Medial superior olive* - olive supérieure médiale

OT *Optic tectum* - tectum optique

PCA *Principal componant analysis* - analyse en composantes principales

RILD *Regularized ILD* - ILD régularisée

RT60 *Reverberation time at 60 dB* - temps de réverbération à 60 dB

SC *Superior colliculus* - colliculus supérieur

SNR *Signal to noise ratio* - rapport signal sur bruit

SOC *Superior olivary complex* - complexe olivaire supérieur

SOM *Self-organizing map* - carte auto-organisatrice

VCN *Ventral cochlear nucleus* - noyau cochléaire ventral

Chapitre 1

Introduction

Sommaire

1.1 Contexte

Aux premières années de la robotique, les robots effectuaient bien souvent une tâche précise et répétitive dans un environnement contrôlé. Les architectures de contrôle employées dans ce contexte ont hérité de l'approche symbolique développée aux origines de l'intelligence artificielle. Cette approche, qui assimile l'intelligence et la perception à une succession de traitements passifs manipulant des symboles de plus en plus abstraits, a permis d'importantes avancées, jouer aux échecs par exemple. L'approche symbolique n'est cependant pas adaptée aux applications actuelles qui demandent à évoluer dans des environnements toujours plus complexes et changeants. Le système « intelligent » ayant battu le champion du monde d'échecs était ainsi incapable de déplacer les pièces sur l'échiquier ou d'inviter son adversaire à une revanche. Alors que ces tâches sont évidentes et immédiates pour un être humain, elles apparaissent à l'inverse très complexes à mettre en oeuvre sur une machine. Ces difficultés rencontrées par l'approche symbolique sont ainsi résumées par le paradoxe de Moravec (1988) : les capacités cognitives de haut-niveau de l'adulte sont plus faciles à modéliser que les capacités sensorimotrices élémentaires du nourrisson. Ces limitations ont incité à ne plus considérer l'intelligence comme une faculté à inférer dans l'abstrait mais comme la capacité à s'adapter à son environnement et répondre à ses besoins internes. Cette approche, résumée par le terme d'intelligence située, s'intéresse davantage aux comportements adaptatifs des animaux qu'aux aspects cognitifs propres à l'intelligence humaine (Guillot & Meyer, 2001; Webb, 2001; Meyer *et al.* , 2005).

La perception est un aspect fondamental de la robotique. Cependant les modèles proposés dans ce domaine reposent bien souvent sur une approche symbolique et, s'ils permettent de bons résultats dans des conditions déterminées, il leur manque généralement une faculté d'adaptation nécessaire à leur utilisation dans des environnements non contrôlés. L'approche symbolique considère donc la perception comme une succession de traitements passifs, dont l'action constitue le résultat final. Dans le

contexte de l'intelligence située, la théorie des contingences sensorimotrices propose un nouveau paradigme pour l'étude de la perception, en plaçant l'action au coeur du processus de perception (O'Regan & Noë, 2001; Philipona *et al.* , 2003). Les capacités de perception reposent alors sur l'analyse du flux sensorimoteur, c'est-à-dire sur les conséquences sensorielles de nos propres actions, pour en extraire des invariants – ou contingences – qui sont caractéristiques des capacités d'interaction d'un agent avec son environnement. La connaissance de ces contingences n'est pas innée mais au contraire découverte par l'expérience sensorimotrice de l'agent.

La vision est historiquement la modalité sensorielle la plus développée dans le domaine de la robotique. Comparativement l'audition est un sens encore récent, le terme *audition robotique* fut vraisemblablement introduit dans la littérature au début des années 2000 (Nakadai *et al.* , 2000a). La modalité auditive présente en effet plusieurs intérêts dans un contexte robotique. Il s'agit de fait d'un sens social prépondérant chez l'homme et le traitement de la parole revêt une importance majeure dans les interactions homme-robot. Le sens de l'audition est également un complémentaire naturel de la vision permettant de détecter des évènements en dehors du champ visuel et d'y diriger l'attention. La localisation de source sonore est ainsi une faculté primordiale de la modalité auditive, sur laquelle repose nombre de capacités de plus haut-niveau. La localisation est également la première application proposée par les modèles d'audition artificielle.

1.2 Objectifs

Cette thèse s'inscrit dans le changement de paradigme proposé par l'intelligence située et la théorie des contingences sensorimotrices. L'objectif de nos travaux est précisément d'appliquer l'approche sensorimotrice à la localisation de sources sonores en robotique. Nous considérons ainsi un robot équipé de capacités motrices élémentaires et d'un système auditif permettant d'extraire des caractéristiques pertinentes pour la tâche de localisation à partir du signal acoustique capturé par une paire de microphones. Ce robot ne possède initialement aucune connaissance ni sur l'environnement ni sur son propre corps et accède à ces informations par l'intermédiaire de son flux sensorimoteur. La question qui nous intéressera est donc la suivante : *Un robot peut-il apprendre la tâche de localisation de sources sonores à partir de l'analyse de son flux sensorimoteur ?* La résolution de cette problématique ouvre en fait sur deux problèmes distincts. D'une part la mise au point d'un système auditif artificiel adapté à la localisation de sources sonores en robotique et d'autre part l'utilisation de ce modèle pour la localisation dans un contexte sensorimoteur.

Ces travaux ont fait l'objet d'une convention CIFRE entre l'entreprise BVS [1] et l'UPMC. Du point de vue applicatif, notre objectif est de développer une approche alternative aux approches symboliques et passives de la perception. La théorie sensorimotrice est encore récente et se base essentiellement sur des arguments philosophiques et psychologiques. Sa « percée » dans le domaine de la robotique est en cours, cette thèse a donc également pour objectif de proposer une avancée dans cette voie.

1. Brain Vision Systems, *www.bvs-tech.com*

1.3 Plan de la thèse

Le chapitre 2 présente les principes biologiques de la localisation de sources sonores chez l'humain et, plus généralement, chez le mammifère. Un intérêt particulier sera porté aux aspects actifs de la perception auditive, et nous verrons à ce titre qu'une approche purement passive ou *bottom-up* de la tâche de localisation n'est pas soutenable. Ce chapitre fournira également les arguments biologiques et psychologiques sur lesquelles reposent les modèles des chapitres 4 et suivants.

Le chapitre 3 propose un état de l'art des méthodes de localisation de sources sonores en robotique en se focalisant sur la « plus-value » que peut apporter le contexte robotique : les méthodes actives d'une part, et les méthodes d'apprentissage d'autre part. Nous verrons ainsi qu'il existe une différence fondamentale entre la grande majorité de ces méthodes robotiques, qui héritent bien souvent d'une conception passive de la perception, et l'organisation fondamentalement active du système auditif biologique. Ce constat motivera l'approche sensorimotrice développée aux chapitres 5 et suivants.

Le chapitre 4 détaille le modèle d'audition binaurale mis au point durant cette thèse. Ce modèle, qui permet d'extraire les indices de base nécessaires à la localisation de sources sonores à partir d'un capteur binaural, se place à mi-chemin entre les modèles neurocomputationnels, visant à modéliser et comprendre les mécanismes biologiques sous-jacents à la localisation, et les modèles « ingénieur-centrés » recherchant avant tout performance et efficacité dans un cadre applicatif bien défini. Ce modèle sera utilisé de manière active et « incarnée » aux chapitres suivants.

Les chapitres 5, 6, 7 et 8 forment un tout cohérent qui applique les concepts sensorimoteurs à la localisation de sources sonores. Le chapitre 5 introduit tout d'abord la théorie sensorimotrice et propose une définition sensorimotrice de la localisation, avant de proposer une première méthode de localisation passive et supervisée. Le chapitre 6 propose une modélisation de comportements de perception active qui permettent une forme primitive de localisation, indissociable du déplacement du robot. En fusionnant les modèles des 2 chapitres précédents, le chapitre 7 propose une méthode d'apprentissage autonome de la localisation : l'expérience sensorimotrice fournie par les capacités de perception active permettent au robot de construire une représentation de son espace sensorimoteur, qui est ensuite utilisée pour une localisation passive. Enfin le chapitre 8 introduit le concept d'ambiguïté perceptuelle, qui permet à l'agent de quantifier de manière totalement autonome le degré de confiance à accorder à une estimation de localisation, et de la corriger de manière active si nécessaire.

Chapitre 2

Principes biologiques de l'audition spatiale

Sommaire

L'audition est une modalité sensorielle sensible aux vibrations transmises par l'air ou par tout autre fluide ou solide. Elle offre aux espèces animales qui en sont équipées la capacité de pouvoir accéder à la fois à la signification d'un son, c'est-à-dire à en interpréter son contenu, et à sa localisation dans l'espace (Kohlrausch *et al.* , 2013). En plus de fournir des informations sur l'environnement, l'analyse et l'interprétation sémantique d'un signal sonore sont directement liées chez de nombreuses espèces aux vocalisations et à la communication intra-espèce (Kanwal & Ehret, 2006), l'audition étant ainsi le sens social prépondérant chez l'humain. De plus, comme l'audition permet une perception au-delà des limites de la vision, sens qui prédomine chez le primate, l'ouïe à également fonction d'alerte et de redirection de l'attention. Concernant la localisation de sources sonores, Bregman (1999) propose une analogie très intuitive pour en introduire le sujet. Imaginons ainsi un lac avec, disposés au creux

d'une de ses rives, deux étroits canaux parallèles entre eux. Au fond de chacun de ces canaux flotte à la surface de l'eau un nénuphar. Amusons-nous un instant à faire quelques ricochets et observons l'onde de choc se propager sur l'eau jusqu'à atteindre les nénuphars. À partir de la seule observation de l'oscillation des nénuphars, peut-on connaître le point de chute d'une pierre ? Peut-on reconstituer la trajectoire effectuée par les ricochets ? Cette courte métaphore illustre évidemment les défis auxquelles est directement confronté le système auditif, le lac représentant l'environnement acoustique, les pierres concrétisant les sources sonores et les nénuphars jouant le rôle des tympans logés au fond des conduits auditifs. Comme nous le verrons au fil de ce chapitre, le système auditif procède à une reconstruction de l'information spatiale grâce à une analyse fréquentielle et temporelle extrêmement fine et complexe, mais également grâce à de nombreux processus *actifs*, c'est-à-dire mettant en oeuvre des mouvements volontaires de l'auditeur. Plus largement le terme *actif* peut également recouvrir des processus adaptatifs non-moteurs.

Il existe une très large littérature dédiée à la biologie du système auditif en général, et à la localisation de sources sonores en particulier. De nombreux ouvrages de référence proposent un état de l'art complet de ce domaine, tant du point de vue psychophysique que neurobiologique (Blauert, 1997; Hartmann, 1997; Moore, 1997; Oertel *et al.* , 2001; Kanwal & Ehret, 2006; Warren, 2008; Yost *et al.* , 2008). L'objectif de ce chapitre n'est bien sur pas de proposer une autre revue de cette littérature mais plutôt d'en proposer une vue synthétique et de poser ainsi le contexte et les connaissances nécessaires à la pleine compréhension de la suite de cette thèse. Nous verrons ainsi que les remarquables performances obtenues par le système auditif en termes de localisation de sources sonores sont rendues possibles grâce à des stratégies variées mettant en oeuvre de nombreux processus actifs et adaptatifs, qui tiennent un rôle fondamental dans le processus de localisation. Ces processus constitueront le "fil rouge" de ce chapitre et nous verrons que le terme *actif* peut se référer à des phénomènes très variés, tant par leurs degrés de complexité que par les niveaux biologiques impliqués, du système auditif central aux zones les plus périphériques. Au-delà du rôle prépondérant de l'action, nous verrons également que l'apprentissage tient une place importante dans le processus de localisation, et notamment qu'une représentation de l'espace auditif, acquise par l'expérience, est exploitée durant le processus de localisation.

Ce chapitre s'intéresse tout d'abord aux capacités de localisation chez l'être humain, à travers le prisme de la psychoacoustique. Le système auditif humain est ensuite décrit en partant de sa périphérie, l'oreille, pour aller vers son centre, le cortex. Les différents éléments anatomiques et neuronaux participant au processus de localisation de sources sonores seront alors introduits.

2.1 Psychologie de l'audition

De nombreuses études de psychoacoustique ont permis depuis le début du XX$^{\text{ème}}$ siècle de précisément décrire les capacités des animaux à l'audition spatiale. Ce paragraphe, en résumant les connaissances majeures de ce domaine, pose les définitions nécessaires et offre un panorama des stratégies adoptées par le système auditif pour parvenir à la localisation d'un son. Les performances de localisation chez l'humain sont tout d'abord introduites dans un contexte purement passif, ce qui nous mènera naturellement à la description des indices auditifs dédiés à cette tâche. Nous verrons ensuite comment le système auditif peut faire face à des environnements complexes,

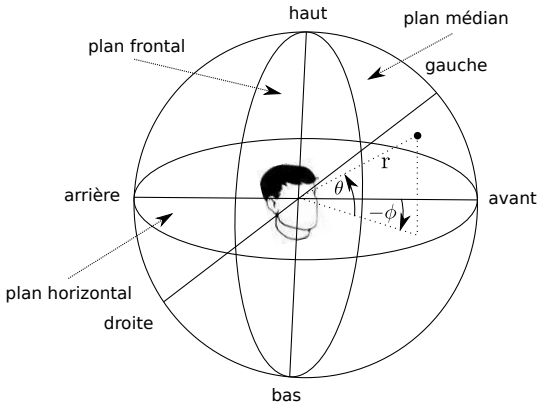


Figure 2.1 – Système de coordonnées relatif à la tête dans lequel est exprimée la position d'une source sonore en fonction de sa distance r, de son angle d'azimut ϕ et de son angle d'élévation θ. Adapté de Blauert (1997).

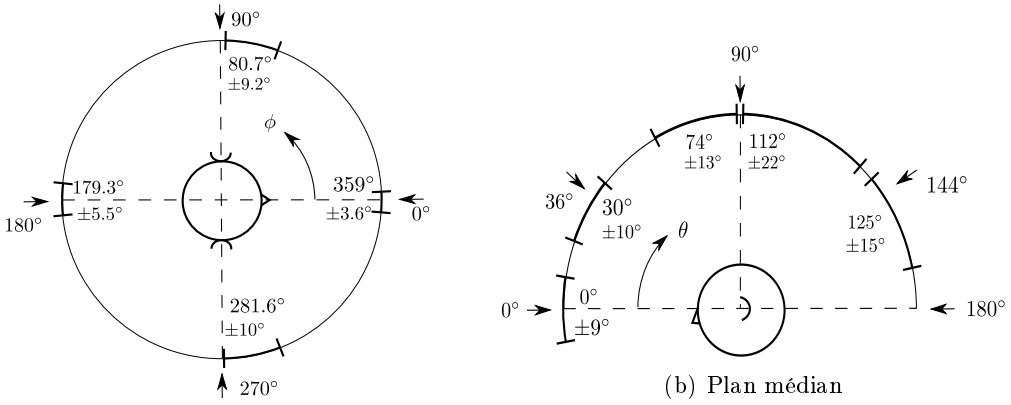
où la présence de réverbération et sources sonores multiples constitue une difficulté majeure. Nous introduirons enfin les différentes stratégies actives employées par les animaux dans le cadre de la localisation de source sonore.

2.1.1 Psychoacoustique

La position d'un son dans l'espace est généralement exprimée dans un système de coordonnées polaires centré sur la tête, comme l'illustre la Fig. 2.1. Ainsi 3 coordonnées doivent être estimées pour une localisation complète : l'angle azimutal ϕ, l'angle d'élévation θ et la distance r de la tête à la source sonore. La majorité des études concernant la localisation considèrent ces 3 coordonnées séparément, et considèrent généralement des conditions statiques, où l'auditeur et la source demeurent immobiles durant tout le temps de l'expérience.

2.1.1.1 Performances de localisation

De par le nombre des expériences proposées et la variété des protocoles expérimentaux utilisés dans la littérature, il est difficile d'établir les performances moyennes de localisation chez l'humain. Cette performance dépend de plus étroitement du signal considéré, de la position de la source et de l'environnement acoustique. Ainsi la résolution angulaire optimale, de l'ordre de 1°, est atteinte pour une source sonore placée directement en face de l'auditeur, à élévation nulle (Blauert, 1997). Néanmoins, à mesure que la source sonore s'éloigne de cette position "idéale", la perception devient de moins en moins précise. Les psychoacousticiens établissent cette perte de précision en mesurant le flou de localisation, défini comme l'angle minimal de déplacement de la source engendrant une modification de sa position perçue par l'auditeur. Ainsi le flou de localisation augmente jusqu'à atteindre son maximum après une rotation de 90° en azimut et de 180° en élévation. Notons également que la localisation est moins précise en élévation, comparativement à la localisation en azimut, et notamment lorsque la source se situe au-dessus ou dans le dos de l'auditeur. La Fig. 2.2 résume les performances de localisation chez l'humain et, si cette localisation est bonne lorsque les angles d'azimut et d'élévation sont faibles, on observe que ces performances se dégradent dans les directions périphériques, avec une sous-estimation systématique de l'angle d'incidence de la source. Enfin, l'estimation



(a) Plan horizontal

(b) Plan médian

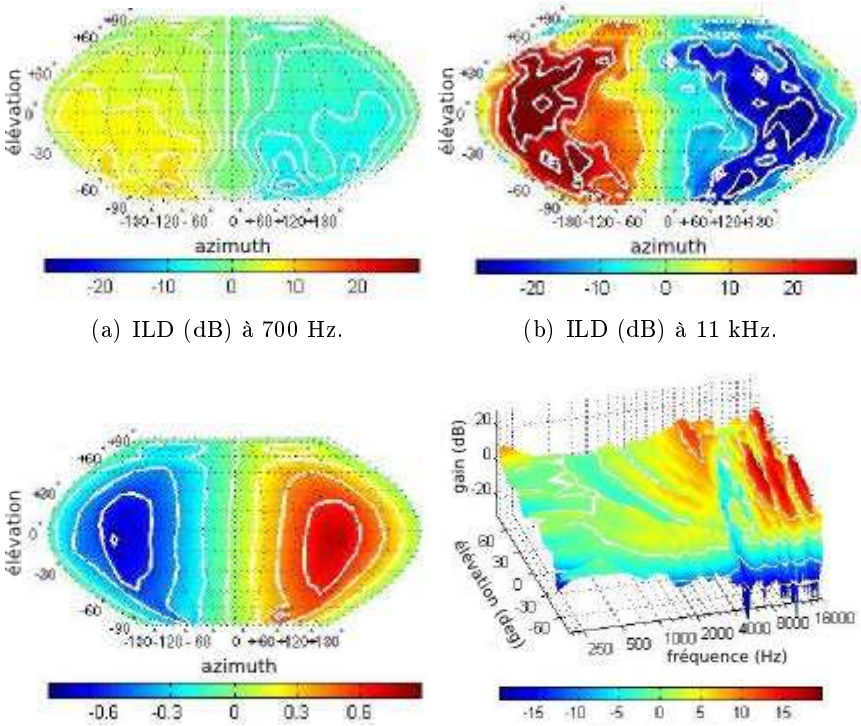
FIGURE 2.2 – Localisation et flou de localisation chez l'humain. (a) Dans le plan horizontal, bruit blanc de 100 ms, 600-900 sujets, d'après Preibisch-Effenberger (1966) et Haustein & Schrimer (1970). (b) Dans le plan médian, voix familière, 7 sujets, d'après Damaske & Wagener (1969). Adapté de Blauert (1997).

de la distance d'une source sonore, décrite au paragraphe 2.1.3, est elle aussi sujette à une importante sous-estimation. Notons finalement qu'une distance importante de la source impacte négativement l'estimation de sa direction, notamment en conditions réverbérantes (Devore *et al.* , 2007).

Ainsi nous constatons que les performances de localisation chez l'homme sont bien loin de la perfection, une erreur d'estimation étant systématiquement rapportée dans les directions périphériques. D'un point de vue écologique, il n'est en effet pas utile de connaître la position d'une source au dixième de degré, seule importe la capacité à pouvoir y réagir : la fuir, s'en approcher, y focaliser son attention, regarder dans sa direction, etc. Ce constat est également à mettre en contraste avec l'objectif de précision motivant bien souvent les modèles de localisation artificiels, comme nous le verrons au chapitre 3.

2.1.1.2 Indices auditifs pour la localisation

Pour parvenir à une estimation de la position d'une source sonore, le système auditif se base sur le calcul de différents indices. Ceux-ci peuvent être monauraux, c'est-à-dire calculés de manière indépendante à gauche et à droite, ou bien binauraux, c'est-à-dire procédant à une mise en correspondance des informations fournies par les deux oreilles. La littérature présente ainsi deux indices binauraux comme fondamentaux pour la localisation dans le plan horizontal (Piéron, 1922; Blauert, 1997). Le premier de ces indices est la différence interaurale d'intensité (ILD), qui est la différence entre l'intensité acoustique aux oreilles gauche et droite (voir Fig. 2.3(a) et 2.3(b)). Le second est quant à lui la différence interaurale de temps (ITD), ou différence interaurale de phase (IPD). Cet indice correspond à la différence entre les temps d'arrivée du signal entre l'oreille gauche et l'oreille droite (voir Fig. 2.3(c)). L'ILD et l'ITD sont les deux composantes de la théorie duplex proposée par Rayleigh (1907), théorie largement confortée par l'expérience, qui stipule que l'ILD est davantage pertinent dans les hautes fréquences, pour lesquels l'ombre acoustique causée par la tête devient prépondérante, et que l'ITD est à l'inverse plus adapté aux basses fréquences, cet indice devenant ambigu dans les hautes fréquences (Blauert, 1997). La



(a) ILD (dB) à 700 Hz.

(b) ILD (dB) à 11 kHz.

(c) ITD (ms).

(d) Indices spectraux (dB).

Figure 2.3 – Indices auditifs mesurés au niveau du tympan sur un sujet humain. (a) ILD en fonction de la direction d'un ton pur à 700 Hz. (b) ILD en fonction de la direction d'un ton pur à 11 kHz. (c) ITD en fonction de la direction de la source sonore. Les lignes d'iso-ITD sont indiquées par les lignes blanches. (d) Indices spectraux monauraux. Du bruit est présenté en face de l'auditeur à différentes élévations. Adapté de King *et al.* (2001).

fréquence séparant les domaines de l'ILD et de l'ITD, dépendante à la fois du signal et de la morphologie de l'oreille externe, se situe autour de 2.5 kHz chez l'humain. La théorie duplex a été vérifiée notamment par Macpherson & Middlebrooks (2002) pour des tons purs. Pour des signaux à large bande, cette dichotomie semble moins marquée et l'ITD domine alors le jugement. Dans l'espace tridimensionnel l'ITD ne permet pas l'estimation de la position absolue d'une source sonore mais estime en fait une surface d'iso-ITD, également appelé cône de confusion (Shinn-Cunningham *et al.* , 2000), l'intersection de deux cônes permettant de lever toute ambiguïté de position.

Les déformations spectrales induites par les réflexions d'une onde sonore sur l'oreille et sur le haut du corps sont également connues pour leur contribution à la localisation de sources dans les plans médian et horizontal (Searle *et al.* , 1975; Blauert, 1997). Ces indices spectraux, principalement monauraux, sont en effet dépendant de la direction de la source sonore (voir Fig. 2.3(d) et paragraphe 2.2.1). Par ailleurs les performances de localisation dans le plan vertical semblent meilleures pour une source à large bande que pour une source au spectre étroit (Roffler & Butler, 1968; Musicant & Butler, 1984). Au-delà de ces indices purement auditifs, il faut également prendre en compte d'autres fonctions cognitives comme la mémoire et l'attention (Palmer *et al.* , 2007), ou encore les informations provenant d'autres

modalités sensorielles, notamment la vision (King, 2009) ou les modalités proprioceptives et somatosensorielles. Pour conclure notre propos sur les indices acoustiques, la Fig. 2.3 nous montre que ces indices peuvent évoluer de manière complexe et non-linéaire en fonction des angles d'incidence de la source, rendant le problème de localisation plus difficile. Cette non-linéarité se vérifie particulièrement sur les Fig. 2.3(b) et 2.3(d), correspondant respectivement à l'ILD en haute fréquence et aux indices spectraux.

2.1.2 Adaptation à l'environnement

Les résultats présentés dans la section précédente se basent sur des conditions expérimentales favorables et sont la plupart du temps obtenus passivement, en conditions anéchoïques et en l'absence de bruit. Cependant les bruits, sources sonores multiples et réverbérations sont omniprésents dans nos environnements quotidiens et le système auditif a développé différentes stratégies pour faire face à ces conditions adverses. Ce paragraphe décrit deux effets psychoacoustiques bien connus qui illustrent le fonctionnement du système auditif dans ces conditions : l'effet de précédence, permettant l'écoute en conditions réverbérantes, et l'effet « cocktail party » regroupant un ensemble de processus permettant d'isoler une source sonore unique confondue dans un environnement acoustique complexe. Ainsi ce paragraphe sera l'occasion d'illustrer le premier aspect de la perception active comme identifié dans l'introduction de ce chapitre, à savoir les processus attentionnels et adaptatifs.

2.1.2.1 Effet de précédence

L'effet de précédence, également appelé loi du premier front d'onde, est un effet psychoacoustique permettant d'augmenter la robustesse de la perception auditive dans des conditions réverbérantes (Wallach *et al.* , 1949). Lorsqu'un son émis à une position donnée est suivi par ce même son émis depuis une autre position avec un retard suffisamment bref, un auditeur ne perçoit pas ces sons comme deux sources distinctes mais comme un percept auditif unique. La position perçue est alors dominée par la direction du premier front d'onde, c'est-à-dire du premier son émis. Cet effet est particulièrement important dans les environnements réverbérants car il permet de percevoir la position d'une source sonore malgré la présence de réflections multiples.

L'effet de précédence peut se décomposer en trois phases distinctes selon le délai qui sépare l'émission des deux sons (Blauert, 1997; Litovsky *et al.* , 1999) :

1. La fusion. Lorsque le délai séparant les deux signaux est de l'ordre de 1 ms environ une source unique est perçue dans une direction intermédiaire entre la position des deux sources.
2. La dominance. À partir de 1 ms et jusqu'à un retard maximal de l'ordre de 35 ms, couramment appelé seuil d'écho, une source unique est perçue dans la direction du premier son, de manière moins nette cependant qu'en l'absence totale d'écho.
3. La discrimination/suppression. Lorsque le retard est supérieur au seuil d'écho deux sources distinctes sont perçues dans les directions correspondant aux angles d'incidence des deux sources. L'écho n'est alors plus perceptible.

Les propriétés précises de ces 3 phases, et notamment la valeur du seuil d'écho, dépendent largement du type du signal étudié. L'effet de précédence est de plus

dépendant du contexte environnemental et met un certain temps à se mettre en place, signe qu'un processus d'adaptation ou d'apprentissage est à l'oeuvre (Litovsky *et al.* , 1999). Enfin, bien que les réflexions ne sont pas perçues par l'auditeur, il demeure tout de même une information permettant de caractériser l'environnement acoustique ambiant, perceptuellement associée à la notion de timbre. Nous proposons dans l'annexe A une simulation reproduisant qualitativement les 3 phases de cet effet de précédence à partir du modèle binaural présenté au chapitre 4.

2.1.2.2 Effet « cocktail party »

L'effet « cocktail party » fut désigné ainsi par Cherry (1953) dans un article devenu classique dans lequel l'auteur s'intéresse à la séparation de deux signaux de parole simultanés. Cet effet, qui regroupe un ensemble de processus permettant d'isoler une source unique dans un environnement multi-source, englobe deux tâches interdépendantes que sont la séparation des différentes sources d'une part, et la capacité à focaliser son attention sur une source particulière d'autre part (Bronkhorst, 2000; McDermott, 2009).

La séparation de sources sonores consiste à séparer le signal auditif en différents flux, chacun de ces flux étant associé à une source distincte. Cette séparation peut s'effectuer sur la base de critères temporels (le changement d'amplitude d'une source se fait généralement de manière simultanée sur l'ensemble de son spectre), fréquentiels (différentes fréquences formant un complexe harmonique appartiennent vraisemblablement à la même source) et spatiaux (des valeurs d'ITD ou d'ILD différents indiquent deux sources distinctes). Ce processus de séparation de sources, ou organisation de scène auditive, est un sujet de recherche à part entière (Bidet-Caulet & Bertrand, 2009; Elhilali *et al.* , 2009; Shamma & Micheyl, 2010; Shamma *et al.* , 2011; McDermott *et al.* , 2011).

A ces processus de séparation s'ajoutent également le rôle des processus efférents et multimodaux. Ainsi le rôle de l'attention auditive à une direction particulière s'avère fondamental (Kidd Jr *et al.* , 2005), de même que le rôle de l'action (Cooke *et al.* , 2007) ou de la vision (Chen, 2003). De par la complexité de ce phénomène et sa nature multidisciplinaire, l'effet « cocktail party » s'avère aujourd'hui encore largement incompris et sa modélisation parcellaire (Chen, 2003).

2.1.3 Audition active

L'information générée par le mouvement du corps, en particulier par les mouvements de la tête et, chez certaines espèces, de l'oreille externe, peut être exploitée par le processus de localisation (Blauert, 1997; Cooke *et al.* , 2007). Les indices acoustiques analysés par le système auditif dépendent en effet de la position de la source par rapport à l'axe binaural et un changement dans la configuration de cet axe engendre une modification de ces indices. Ce paragraphe présente donc les principaux résultats concernant la localisation active de sources sonores et le terme *actif* s'entend ici au sens traditionnel, c'est-à-dire impliquant directement un mouvement volontaire ou non de la part de l'auditeur. Nous présentons d'abord quelques études concernant le mouvement de la tête chez l'humain et le chat puis nous nous intéresserons au cas particulier de la perception de la distance, qui illustre de façon tout à fait singulière la complexité des procédés mis en oeuvre durant ces comportements d'audition active.

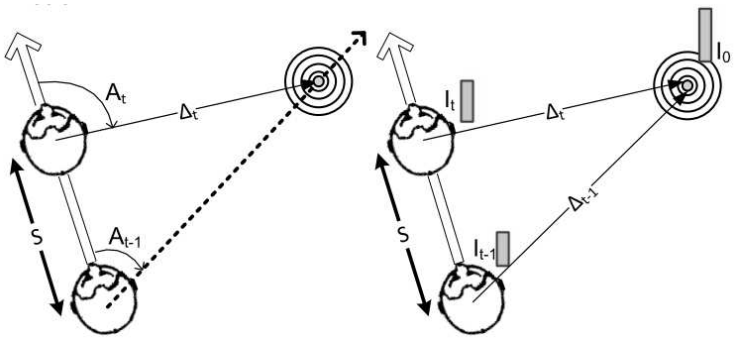


FIGURE 2.4 – Parallaxe de mouvement et τ acoustique pour l'estimation active de la distance durant un mouvement de translation de l'auditeur. La parallaxe de mouvement (gauche) estime le changement d'azimut induit par la translation tandis que le τ acoustique (droite) évalue le changement d'intensité. D'après Lu *et al.* (2007)

2.1.3.1 Mouvements de la tête

Dans sa revue bibliographique consacrée aux comportements de localisation exploitant les mouvements de la tête, Blauert (1997) indique que l'objectif principal de ces comportements est d'obtenir une information plus précise sur la localisation de la source mais qu'une estimation même grossière de cette position est nécessaire avant tout mouvement. Ainsi, des mouvements légers et spontanés de la tête sont utilisés pour lever certaines ambiguïtés, des mouvements plus importants permettent d'améliorer le rapport signal sur bruit (SNR), de diminuer le flou de localisation ou encore de minimiser les interférences avec d'autres sources (Young, 1931; Wallach, 1940; Loomis *et al.* , 1990). Une brève rotation en aller-retour peut également être observée. Enfin l'audition active peut également participer à la localisation en élévation (Wallach, 1940; Perrett & Noble, 1997). Certains mouvements de recherche volontaire s'inscrivent dans un cadre plus complexe, dans lequel l'information est intégrée et assemblée en continu durant l'action. Ce comportement s'illustre notamment lors du suivi ou de l'approche d'une source en mouvement (Loomis *et al.* , 1990), ou encore lors de l'estimation de la distance comme nous le verrons ci-dessous. Un réflexe d'orientation de la tête vers une source sonore est également présent chez le nouveau-né et constitue d'ailleurs le premier comportement visible associé à l'audition spatiale (Kearsley, 1973). Ce réflexe semble prendre moins d'importance à mesure que le nourrisson grandit et semble également prendre part à l'apprentissage subséquent de la localisation de sources sonores (Muir *et al.* , 1989). Ces mouvements d'orientation et de suivi, dans le cas d'une source sonore mobile, sont présents chez l'homme mais ont également été observés chez d'autres espèces animales comme la chouette effraie, la chauve-souris, la gerbille et le chat (Kelly & Potash, 1986; Beitel, 1999). Ainsi le chat exploite la mobilité de ses pavillons pour la localisation de sources sonores (Populin & Yin, 1998), avec un mouvement en deux phases. À court terme (25 ms), les pavillons s'orientent grossièrement vers la direction de la source, tandis qu'à plus long terme ceux-ci ont tendance à accompagner le mouvement des yeux, bien souvent suivis d'un mouvement d'orientation de la tête (Beitel, 1999).

2.1.3.2 Perception de la distance

La distance est généralement estimée avec une large erreur, notamment pour un éloignement important. Comme dans le cas de l'estimation de la direction d'une

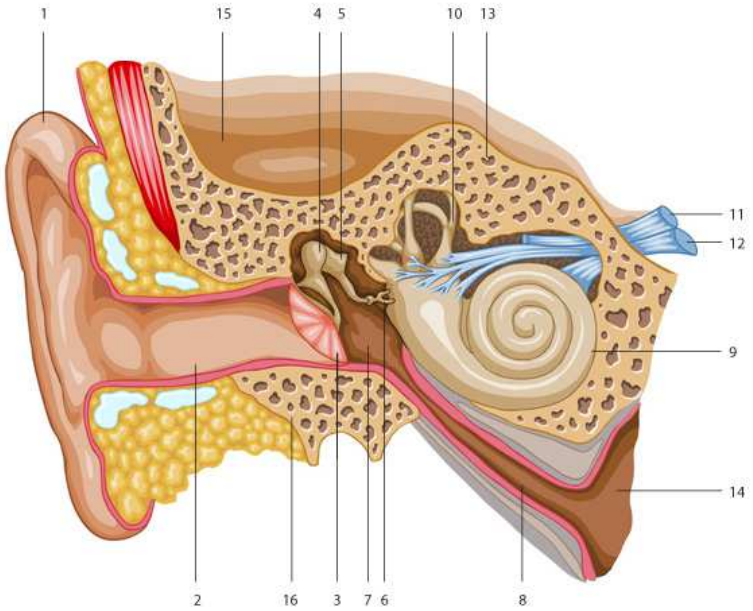


Figure 2.5 – Anatomie du système auditif périphérique comprenant les oreilles externe (1-2), moyenne (3-7) et interne (8-12). 1 - pavillon. 2 - conduit auditif externe. 3 - tympan. 4 - marteau. 5 - enclume. 6 - étrier. 7 - caisse du tympan. 8 - trompe d'Eustache. 9 - cochlée. 10 - système vestibulaire. 11 - nerf vestibulaire. 12 - nerf auditif. 13 - rocher. 14 - fosse nasale. 15 - cerveau. 16 - mastoïde. Adapté de Stangor (2010).

source sonore, il existe pour l'estimation de la distance un biais psychophysique important entraînant une sous-estimation systématique de la distance perçue par rapport à la distance réelle (von Békésy, 1949; Loomis *et al.* , 1990; Blauert, 1997). La perception de la distance est en effet un problème complexe mettant en oeuvre différents indices monauraux et binauraux simultanément (Zahorik, 1996, 2002; Naguib, 2001; Cooke *et al.* , 2007), nécessitant par ailleurs une prise en compte de la dynamique de ces indices, lorsque l'auditeur ou la source sont mobiles, et des connaissances *a priori* sur la source et sur l'environnement considérés. Ainsi des indices monauraux comme l'intensité, le rapport entre énergie directe et énergie réverbérée ou l'atténuation des hautes fréquences sont utilisés pour l'estimation de la distance d'une source, mais ils ne peuvent pas l'être sans connaissance *a priori* (Mershon & Bowers, 1979). De plus les comportements actifs peuvent être exploités efficacement pour cette tâche. Un simple mouvement en translation permet ainsi de générer un changement dans l'azimut et dans l'intensité perçue et permet d'estimer la distance grâce au calcul de la parallaxe de mouvement et du τ acoustique (Ashmead *et al.* , 1995; Speigle & Loomis, 2002), comme illustré par la Fig. 2.4.

2.2 Système auditif périphérique

Le paragraphe précédent s'est intéressé à la localisation de sources sonores d'un point de vue psychophysique et extérieur. Il est maintenant temps de rentrer "à l'intérieur" du système auditif pour mieux comprendre comment il procède à la localisation d'un son. D'un point de vue anatomique, le système auditif périphérique se compose

de trois parties distinctes que sont l'oreille externe, l'oreille moyenne et l'oreille interne. Les oreilles externe et moyenne ont un rôle de transmission de l'information acoustique de l'environnement extérieur vers l'oreille interne, cette dernière ayant un rôle de réception et de conversion de l'énergie acoustique en influx nerveux. La limite entre systèmes auditifs périphérique et central peut ainsi se situer à la « sortie » de l'oreille interne, c'est-à-dire à l'interface entre les cellules ciliées et le nerf auditif (de Cheveigné, 2004). L'organisation anatomique et fonctionnelle du système auditif périphérique humain, illustré Fig. 2.5, est brièvement décrite dans les paragraphes suivants. Nous verrons à ce titre que les aspects actifs sont là encore présents à tous les niveaux.

2.2.1 Oreille externe

Charles Darwin ne prêtait pas grande utilité au pavillon chez l'homme, pensant qu'il était le vestige d'une structure plus élaborée présente chez nos ancêtres et d'autres espèces, tels que les pavillons du chat ou de la chauve-souris (Warren, 2008). Cette opinion est aujourd'hui largement contredite et il est démontré que l'oreille externe joue au contraire un rôle majeur dans le processus de localisation de sources sonores (Batteau, 1967). Ainsi l'oreille externe est présente chez la totalité des espèces de mammifères terrestres, elle inclut le pavillon et le conduit auditif externe ainsi que, au sens large, la tête et le torse. Sa morphologie induit sur le signal acoustique des transformations nombreuses et complexes, telles que atténuations, amplifications, diffractions et réverbérations dépendant à la fois du contenu fréquentiel du signal, de la position de la source relativement à l'auditeur mais également de la morphologie du pavillon. Le filtrage opéré par l'oreille externe apporte ainsi des informations concernant la position avant-arrière d'une source sonore, ainsi que son azimut et son élévation (Rayleigh, 1907; Batteau, 1967; Rice *et al.* , 1992; Zakarauskas & Cynader, 1993). Son efficacité est accrue notamment pour les sources rapprochées et composées de hautes fréquences (Shaw & Teranishi, 1968; Musicant & Butler, 1984).

Les propriétés du filtrage effectué par l'oreille externe peuvent être représentées par une fonction de transfert relative à la tête (HRTF). La HRTF est définie comme le rapport entre la pression acoustique au niveau du tympan et la pression acoustique à la position du centre de l'axe interaural dans un champ acoustique libre. Nous avons dit que cette HRTF varie en fonction de la fréquence et de la position d'une source. Mais, en conséquence de l'individualité de l'oreille externe, la HRTF varie également d'un sujet à l'autre et, pour le même sujet, évolue également au cours de l'existence. Ainsi cette fonction de transfert nécessite d'être apprise par le système auditif central, cet apprentissage se mettant à jour lorsque le filtrage est modifié (Hofman *et al.* , 1998).

Le conduit auditif externe, intermédiaire entre le pavillon et le tympan, est quant à lui bien plus qu'un simple conduit passif. Il se comporte en effet comme un résonnateur et augmente l'impédance du milieu de plus de 5 dB pour les fréquences comprises entre 2 kHz et 5.5 kHz chez l'humain, bande fréquentielle cruciale pour la compréhension de la parole (Wiener, 1947). Finalement, le changement de pression acoustique à l'extrémité interne du conduit auditif engendre une vibration du tympan.

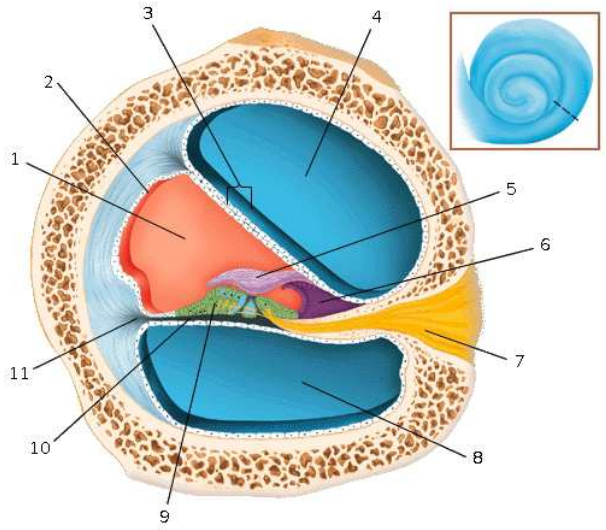


Figure 2.6 – Vue en coupe de la cochlée. 1 - canal cochléaire. 2 - strie vasculaire. 3 - membrane de Reissner. 4 - rampe vestibulaire. 5 - membrane tectorielle. 6 - limbe spiral. 7 - nerf auditif. 8 - rampe tympanique. 9 - organe de Corti. 10 - membrane basilaire. 11 - ligament spiral. D'après *Encyclopedia Britannica*, 1997.

2.2.2 Oreille moyenne

L'oreille moyenne transmet l'énergie acoustique depuis le tympan vers la cochlée au travers d'une chaîne de trois osselets : le marteau en contact direct avec le tympan, l'enclume et l'étrier relié à la cochlée via la fenêtre ovale (voir Fig. 2.5). Le rôle principal de l'oreille moyenne est d'adapter la faible impédance du milieu acoustique en entrée pour sa diffusion dans le milieu liquide de la membrane basilaire, au sein de la cochlée, où l'impédance y est environ 4000 fois plus importante (Fay & Popper, 2010).

L'oreille moyenne agit en première approximation comme un système linéaire pour les intensités sonores inférieurs à 130 dB SPL. Néanmoins des études plus poussées font apparaître des non-linéarités pour les fréquences supérieures à 1 kHz (Sinyor & Laszlo, 1973). De plus, les intensités les plus élevées sont atténuées par le réflexe stapédien, jusqu'à une valeur de 30 dB SPL pour les fréquences en deçà de 1 kHz (Warren, 2008). Ce réflexe est un processus actif d'atténuation du signal sonore qui sert à protéger la cochlée d'amplitudes trop élevées. Il peut à ce titre être comparé au réflexe contractant l'iris de l'oeil en fonction de la luminosité ambiante.

2.2.3 Oreille interne

L'oreille interne est composée de la cochlée, organe de l'audition, et de l'appareil vestibulaire, organe de l'équilibre responsable de la perception de la position angulaire de la tête et de son accélération. C'est au sein de la cochlée, qui tient son nom de la racine latine du mot "escargot", que l'énergie mécanique transmise par l'étrier via la fenêtre ovale est convertie en influx nerveux puis transmise au nerf auditif. Les vibrations acoustiques se propagent ainsi au travers de la fenêtre ovale dans la périlymphe, le milieu liquide contenu dans les 2,5 tours de spire de la cochlée, de la base vers l'apex dans la rampe vestibulaire puis de l'apex vers la base dans la rampe tympanique, pour en sortir via la fenêtre ronde. Ces deux rampes sont

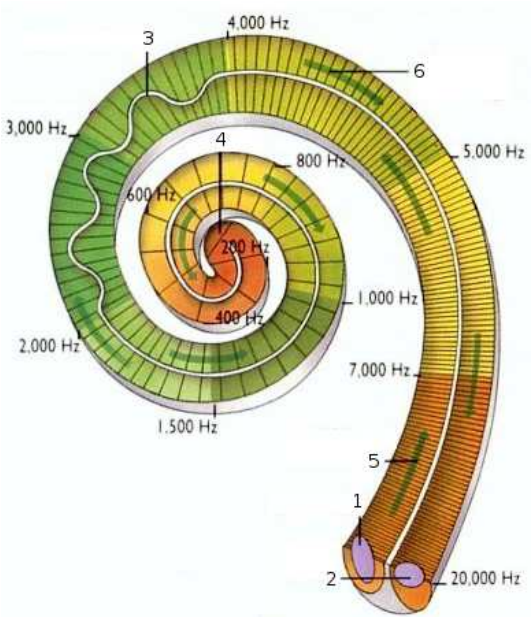


FIGURE 2.7 – Organisation tonotopique d'une cochlée humaine. L'échelle de couleurs représente l'intensité du mouvement de la membrane basilaire pour une excitation dans la bande des 3 kHz. 1 - fenêtre ovale. 2 - fenêtre ronde. 3 - membrane basilaire. 4 - apex (hélicotrème). 5 - énergie entrante (rampe tympanique) . 6 - énergie sortante (rampe vestibulaire). Adapté de *http ://universe-review.ca.*

reliées au niveau de l'apex par l'hélicotrème et sont séparées le long de la cochlée par le canal cochléaire. La membrane basilaire est une membrane souple séparant la rampe tympanique du canal cochléaire, comme illustré Fig. 2.6. Celle-ci a des propriétés complexes revues avec détail par Robles & Ruggero (2001), qui assurent une sensibilité fréquentielle très fine. Ainsi un point donné le long de cette membrane est sensible à une fréquence spécifique. Une décomposition fréquentielle du signal acoustique s'effectue en conséquence le long de cette membrane, depuis les hautes fréquences à la base de la cochlée vers les basses fréquences au niveau de l'apex.

Cette décomposition fréquentielle opérée par la membrane basilaire, illustrée Fig. 2.7, permet d'obtenir une représentation tonotopique de l'information auditive, c'est-à-dire littéralement une représentation spatiale de l'information fréquentielle. Comme nous le verrons dans la suite de ce chapitre, cette tonotopie est une caractéristique fondamentale dans l'organisation du système auditif. Elle est en effet retrouvée à tous les niveaux, de la cochlée jusqu'au cortex (Warren, 2008). La mécanique cochléaire présente enfin des aspects non-linéaires, pour certains toujours mal compris (Warren, 2008), qui permettent entre autre l'amplification des signaux de faible amplitude, la compression de stimuli trop intenses ou encore le démasquage d'une fréquence masquée par une fréquence proche d'intensité plus élevée, notamment grâce une adaptation de la sensibilité fréquentielle de la membrane basilaire (Plack & Oxenham, 1998; Oxenham, 2001; Plack *et al.* , 2002).

A l'intérieur du canal cochléaire, reposant sur la membrane basilaire, se trouve l'organe de Corti (voir Fig. 2.8) dans lequel sont fixées les cellules ciliées, les cellules sensorielles de l'appareil auditif reliées aux fibres du nerf auditif. Il existe deux sortes de cellules ciliées, les cellules ciliées internes ou CCI (Fettiplace & Fuchs, 1999) et les cellules ciliées externe ou CCE (Fettiplace & Hackney, 2006), qui diffèrent par

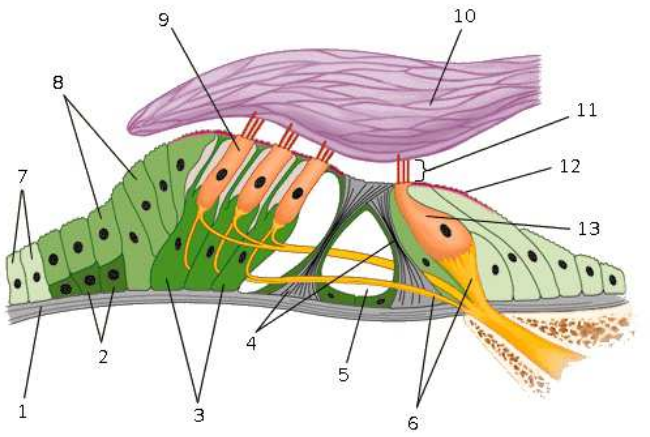


Figure 2.8 – Organe de Corti. 1 - membrane basilaire. 2 - cellules de Boettcher. 3 - cellules de Deiters. 4 - pilier de Corti. 5 - tunnel de Corti. 6 - fibres nerveuses. 7 - cellules de Claudius. 8 - cellules de Hensen. 9 - cellule cillée externe. 10 - membrane tectorielle. 11 - stéréocils. 12 - lamina réticulaire. 13 - cellule ciliée interne. D'après *Encyclopedia Britannica*, 1997.

leur morphologie et leur innervation, et donc par leur fonction. Les CCI ont ainsi un rôle principalement afférent tandis que les CCE ont un rôle efférent (voir paragraphe 2.3.3). Une déformation mécanique de la membrane basilaire, générée par une différence de pression entre les rampes vestibulaire et tympanique, provoque l'ouverture de canaux ioniques à la surface des cils des CCI, impliquant une modification de leur potentiel intracellulaire. Ce changement de potentiel entraîne alors la libération d'un neurotransmetteur qui excite les fibres afférentes innervant les CCI. L'interface entre les systèmes auditifs périphérique et central se situe au niveau de la connexion entre les cellules ciliées et le nerf auditif, cette interface étant donc à la fois afférente, avec l'innervation des CCI, et efférente, avec l'innervation des CCE.

2.3 Bases neurales de la localisation auditive

Le système auditif central peut se décrire en première approximation comme une hiérarchie ascendante de noyaux cérébraux, allant du ganglion spinal jusqu'au cortex auditif (de Cheveigné, 2004). Les fibres afférentes du nerf auditif sont en effet constituées par les axones des neurones du ganglion spiral, le premier noyau du système auditif central. Les principaux niveaux suivants sont le noyau cochléaire (CN), le complexe olivaire supérieur (SOC), les noyaux du léminisque latéral et le colliculus inférieur (IC), tous situés dans le tronc cérébral. Au plus haut niveau se trouvent le corps genouillé médian (MGB), dans le thalamus, et le cortex auditif. Cette organisation hiérarchique du système auditif central, représentée sur la Fig. 2.9, peut également se décomposer en voies auditives primaire et secondaires. La voie primaire, courte (3 à 4 relais) et rapide, se termine dans le cortex auditif primaire. Les voies secondaires sont quant à elles fusionnées avec d'autres modalités dès la sortie de CN et sont impliquées dans des facultés comme l'attention ou les réactions végétatives. Nous l'avons déjà mentionné mais il semble utile de le rappeler ici, l'organisation tonotopique offerte par la cochlée se retrouve à tous les niveaux de traitement du système auditif central.

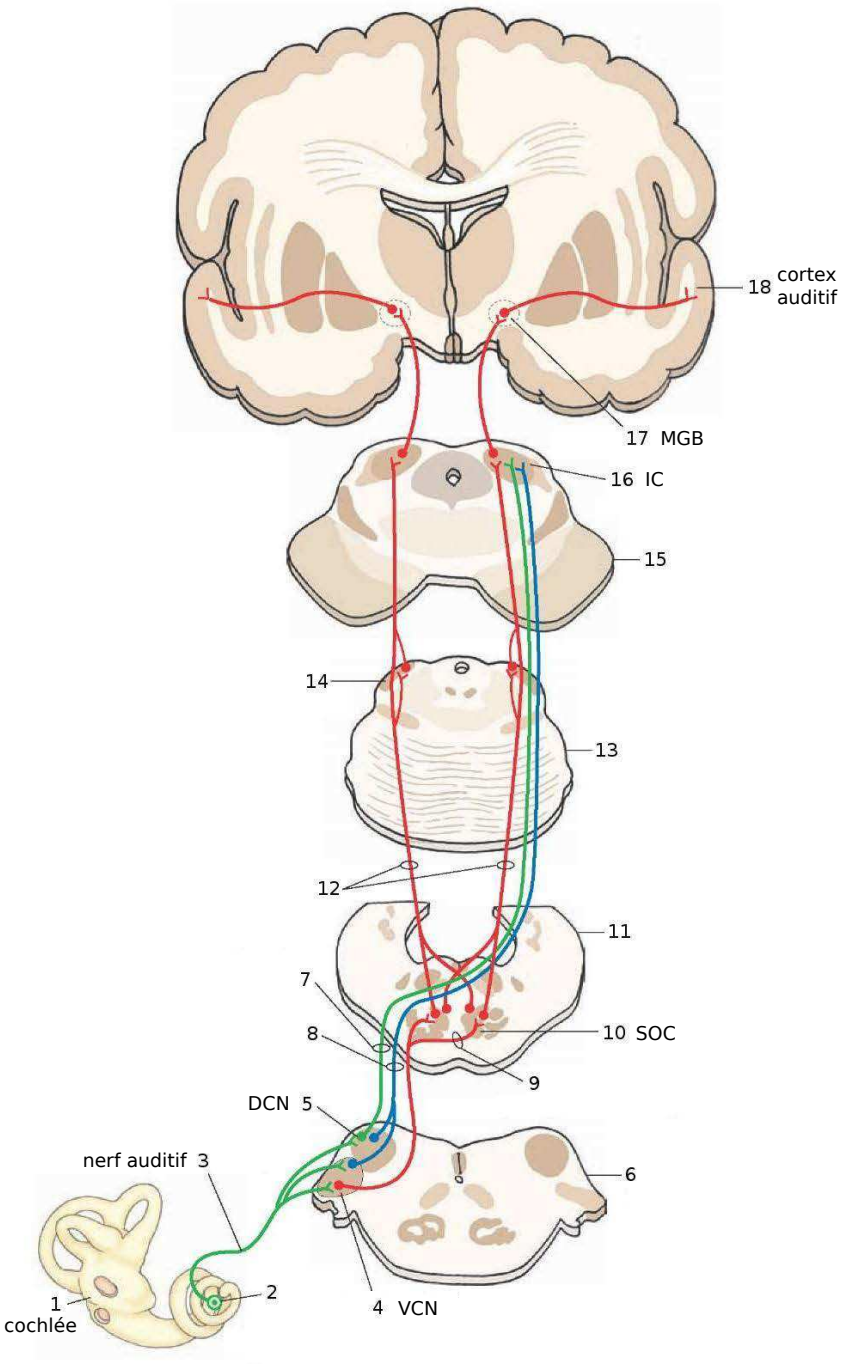


FIGURE 2.9 – Voies auditives primaire (rouge) et secondaires (vert et bleu) du système auditif central chez l'homme. 1 - cochlée. 2 - ganglion spiral. 3 - nerf auditif. 4 - noyau cochléaire ventral (VCN). 5 - noyau cochléaire dorsal (DCN). 6 - bulbe rachidien. 7 - strie acoustique dorsale. 8 - strie acoustique intermédiaire. 9 - corps trapézoïdal (strie acoustique ventrale). 10 - complexe olivaire supérieur (SOC). 11 - pont. 12 - léminisque latéral. 13 - jonction pont-mésencéphale. 14 - noyau du léminisque latéral. 15 - mésencéphale. 16 - colliculus inférieur (IC). 17 - corps genouillé médian (MGB). 18 - cortex auditif. D'après *The Crankshaft Publishing, www.what-when-how.com.*

Ce paragraphe décrit le système auditif central d'un point de vue anatomique et fonctionnel. Les différents noyaux sont introduits suivant la hiérarchie ascendante décrite ci-dessus. Nous nous focaliserons néanmoins sur les traitements auditifs associés à la localisation. Nous introduisons successivement les noyaux du tronc cérébral, le système thalamocortical et enfin le système efférent qui est à l'origine de nombre de processus actifs décrits dans ce chapitre.

2.3.1 Tronc cérébral

CN, SOC et IC sont les principaux noyaux auditifs du tronc cérébral du mammifère. Les détails anatomiques et physiologiques ne seront pas abordés mais peuvent être retrouvés par exemple dans la revue bibliographique proposée par Yin (2002), d'où sont issues la plupart des informations présentées ci-dessous.

CN comprend une division dorsale (DCN) et ventrale (VCN) et reçoit principalement ses entrées du nerf auditif. De part la présence de neurones spécialisés dans le traitement de l'information temporelle, VCN semble être dédié à l'amélioration de la résolution temporelle. DCN quant à lui est spécialisé dans l'extraction des indices spectraux causés par l'oreille externe et contribue ainsi à la localisation en élévation. Il contient en effet des cellules à la sensibilité fréquentielle très complexe et reçoit des projections des systèmes somatosensoriel et vestibulaire, indiquant vraisemblablement une prise en comte de la position ou du déplacement de l'oreille externe (Oertel & Young, 2004).

SOC est le principal site de convergence des sorties de CN. Il se compose des olives supérieures latérale (LSO) et médiale (MSO). Ces noyaux sont connus pour leur contribution à la localisation en azimut. En détectant les différences temporelles entre ses entrées ipsilatérales et contralatérales, MSO est spécialisé dans l'estimation de l'ITD tandis que le LSO, dont les entrées ipsilatérales sont excitatrices et les entrées contralatérales sont inhibitrices, est lui spécialisé dans le calcul de l'ILD. En accord avec la théorie duplex présentée précédemment, le MSO et LSO reçoivent principalement des projections en basses fréquences et hautes fréquences respectivement.

IC est le noyau auditif du tronc cérébral qui comprend le plus de neurones. Il reçoit des projections de la plupart des noyaux inférieurs et projette de manière massive vers le cortex via MGB. IC joue à ce titre le rôle d'interface ou de « gare de triage » entre les noyaux auditifs de niveaux inférieurs et le cortex auditif, tout en permettant de moduler une réponse motrice et comportementale (Casseday *et al.* , 2001). IC procède également à l'intégration des différents indices acoustiques. C'est ainsi dans le noyau central de IC (ICC) que sont fusionnées les données concernant les indices spectraux provenant de DCN et les données binaurales projetées depuis SOC. Puisque IC procède à une intégration des indices spatiaux, on pourrait s'attendre à y trouver une carte neurale de l'espace auditif. Bien qu'une telle carte ait été retrouvée dans le noyau externe du IC (ICX) de la chouette effraie (Bergan & Knudsen, 2008), sa recherche chez le mammifère n'a pas été un complet succès car, bien que des neurones semblent être sensibilisés à la direction d'un son, ils sont également sensible à d'autres dimensions comme le spectre ou les modulations en amplitude ou fréquence notamment, ce qui rend l'enregistrement et l'interprétation de telles cartes hasardeux (Casseday *et al.* , 2001). IC n'en reste pas moins un relais majeur de l'information auditive dans son trajet vers le cortex. Mais IC projette également vers le colliculus supérieur (SC), un noyau spécialisé dans l'intégration multimodale (Stein & Meredith, 1993). Cette projection est particulièrement bien

comprise chez la chouette effraie où la carte auditive de ICX est projetée vers le tectum optique (OT), l'équivalent chez la chouette de notre SC, et alignée sur une carte rétinotopique de l'espace visuel (Bergan & Knudsen, 2008). Cet alignement de l'espace de représentation auditif sur l'espace de représentation visuel n'est pas inné mais est au contraire appris (Bergan & Knudsen, 2008; Werner, 2008). Son apprentissage à principalement lieu lors de l'enfance même si une adaptation est démontrée à l'âge adulte chez plusieurs espèces, notamment pour compenser les changements perceptifs dus à l'âge. Les sorties motrices au niveau de OT chez la chouette ou de SC chez le mammifère sont impliquées dans le contrôle de mouvements d'orientation qui permettent un focus de l'attention sur un évènement spatialement localisé, indépendamment de sa modalité (King *et al.* , 2001), comme le réflexe vestibuloculaire qui permet de stabiliser le champ visuel durant une rotation de la tête, où encore le réflexe de Startle qui permet de réagir à un stimulus brusque et soudain (Lang *et al.* , 1990).

2.3.2 Système thalamocortical

La cible terminale de la voie auditive primaire est bien sûr le cortex auditif, situé dans le lobe temporal (Winer & Schreiner, 2011). Le cortex auditif est en étroite interaction avec la partie auditive du thalamus, le MGB, lui-même étant massivement connecté à ICC (King *et al.* , 2001). Le cortex auditif peut être en première approximation divisé en deux entités. Le cortex auditif primaire, tout d'abord, reçoit principalement ses entrées de la partie ventrale de MGB et présente une forte tonotopie. Il est impliqué dans l'évaluation de critères bas-niveaux des sons perçus, tels que leur hauteur et leur intensité. Le cortex auditif périphérique enfin reçoit ses entrées du cortex primaire et des aires périphériques de MGB. Il exhibe une tonotopie plus diffuse que le cortex auditif primaire, rendant plus complexe la détermination précise de ses fonctions. Le cortex auditif humain est de plus extrêmement sensible à la voix, cette sensibilité pouvant être comparée à celle du cortex visuel pour la détection de visages (Belin *et al.* , 2000).

Concernant plus précisément la localisation de sources sonores, nous venons de voir que les traitements principaux sont effectués au niveau du tronc cérébral par SOC, IC et SC. Néanmoins, il est connu de longue date qu'un dysfonctionnement du cortex auditif entraîne une perturbation de la localisation, indiquant une implication du cortex dans cette tâche (King & Middlebrooks, 2011). Ainsi certaines aires corticales semblent être particulièrement sensibles à la présence de sources multiples (Zatorre *et al.* , 2002). D'autres aires sont quant à elles liées à l'évaluation de la dynamique d'indices binauraux permettant la perception du mouvement (Krumbholz *et al.* , 2005), ou encore à la modulation attentionnelle (Palmer *et al.* , 2007). Ces trois exemples d'implication du cortex auditif dans une tâche de localisation sont directement liés à des comportements actifs, qu'ils soient moteurs ou efférents, puisque des voies descendantes permettent de moduler la réponse des noyaux en amont.

2.3.3 Système efférent

La présence de voies efférentes est un facteur de complexité très important dans la compréhension du sens de l'audition et le système efférent est encore aujourd'hui mal compris (Robles & Delano, 2008). Ce système se caractérise en effet par la présence massive de connexions efférentes à tous les niveaux de traitement. Ces connexions ont pour origine le cortex, où les fibres efférentes sont plus nombreuses que les fibres

afférentes, notamment dans la boucle thalamocorticale. Le système efférent descend alors jusqu'à la cochlée, en passant par les différents noyaux auditifs du tronc cérébral, à l'exception de CN. Les CCE de la cochlée sont en effet directement innervées par des neurones du MSO.

Différents rôles possibles ont été attribués au système efférent par différents auteurs (Robles & Delano, 2008; Guinan, 2010), comme le démasquage de stimuli en présence de bruit, un rôle protecteur contre les sons de trop forte intensité et les pertes auditives temporaires ou permanentes, une modulation de la sensitivité cochléaire en fonction de l'attention, la modulation des réponses afférentes durant le sommeil et plus récemment une modulation de la réponse binaurale dans le LSO, lié au calcul de l'ILD. Malgré ces différentes hypothèses les fonctions précises du système efférent sont encore largement inconnues mais semblent être impliquées dans des aspects variés de l'audition.

La partie la plus périphérique du système efférent, reliant le MSO aux CCE, est sa composante qui est la mieux étudiée et la mieux comprise. Elle a également fait l'objet d'une modélisation (Ferry & Meddis, 2007). Ce modèle, qui permet d'atténuer la réponse de la membrane basilaire en fonction de l'activité de MSO, reproduit des résultats expérimentaux effectués sur l'animal sur des signaux simples. L'objectif des auteurs est d'étudier le rôle fonctionnel du système efférent en présence de stimuli et de conditions acoustiques plus complexes. Ce modèle a ainsi été appliqué à la reconnaissance de parole (Brown *et al.* , 2010) et les auteurs démontrent que le démasquage causé par l'activité efférente permet d'améliorer le taux de reconnaissance de parole bruitée.

2.4 Discussion

Cette discussion aborde tout d'abord quelques facteurs de complexités ignorés dans la description du système auditif présentée dans ce chapitre. Nous revenons ensuite sur la notion de perception active et sur le rôle de l'apprentissage dans le processus de localisation. Enfin nous présentons quelques "imperfections" du système auditif mettant en lumière la nature sensorimotrice de la perception.

Facteurs de complexité La description du système auditif proposée jusqu'ici constitue une approximation sévère de sa réelle complexité. Il apparaît en effet que des facteurs viennent enrichir cette description simple (de Cheveigné, 2004). Il s'agit premièrement de la structure bilatérale du cerveau. Nous l'avons jusqu'ici ignoré mais les différents noyaux du système auditif sont tous présents en deux exemplaires, depuis les oreilles jusqu'au cortex, les différents noyaux centraux étant connectés par un complexe réseau de connexions ipsi et contralatérales. Une symétrie anatomique et fonctionnelle se retrouve pour les niveaux les plus périphériques, jusqu'à IC, tandis que le cortex et le MGB présentent une spécialisation hémisphérique.

Un second facteur de complexité est la présence de connexions transhiérarchiques. Il existe dans la hiérarchie ascendante vers le cortex des connexions reliant un à un chacun des noyaux cérébraux, nous l'avons dit, mais il existe également de très nombreuses connexions transhiérarchiques, c'est-à-dire des connexions court-circuitant un ou plusieurs noyaux dans la chaîne ascendante. CN, IC et MGB sont cependant autant d'étapes imposées par la connectique du système auditif, qui ne peuvent pas être évitées. Il existe également dans le système efférent des connections transhiérarchiques parallèles aux connections descendantes directes à tous les niveaux, à

l'exception de IC qui apparaît là encore comme un routeur de l'information auditive en direction et en provenance du cortex (Oertel *et al.* , 2001).

Perception active La majorité des études concernant le système auditif, comme les modèles généralement proposés en robotique (voir chapitre 3), considèrent l'audition comme un processus purement *bottom-up*. Un tel système est composé de voies afférentes uniquement et les différents centres de traitement sont indépendants les uns des autres (Slaney, 1997). Cette approximation est bien souvent souhaitable et nécessaire, néanmoins ce chapitre a montré l'importance et le rôle prépondérant des processus *top-down*. En témoignent la présence massive de connexions efférentes à tous les niveaux de traitements, ou encore les implications du système moteur à différents degrés de cognition. Il apparaît donc que ces deux aspects *bottom-up* et *top-down* sont intrinsèquement liés et que leur interaction est permanente à tous les niveaux du système auditif.

Ce chapitre a ainsi identifié plusieurs phénomènes actifs associés à la localisation : mouvements de la tête, déplacements vers la source, mouvements des pavillons ou écholocation chez certaines espèces. Grâce aux modèles de localisation active présentés aux chapitres 5 et suivants, nous verrons que la théorie sensorimotrice offre un formalisme tout à fait adapté à la modélisation de ce type de comportements actifs. D'autres phénomènes actifs non-moteurs interviennent également, tels que l'effet cocktail ou le contrôle efférent de la cochlée. La modélisation de ceux-ci ne sera pas abordée. La possibilité d'une modélisation sensorimotrice du système efférent est néanmoins abordée en conclusion de cette thèse.

Apprentissage Nous avons également vu dans ce chapitre que l'apprentissage est un élément important du processus de localisation : apprentissage des HRTF ou l'alignement rétinotopique de l'espace auditif notamment. Cette idée que le système nerveux central « construit » une représentation de l'espace auditif par apprentissage sera à la base du chapitre 5, qui présente une méthode de localisation basée sur une telle représentation.

Constatons enfin que la géométrie de l'espace auditif tel qu'il est représenté par le système nerveux n'est pas isométrique à la géométrie de l'espace acoustique réel. En témoignent les biais systématiques que l'on peut observer dans une tâche de localisation de sources sonores pour un azimut, une élévation ou une distance élevée. On retrouve ce même phénomène en vision, ou le centre du champ visuel, traité par la fovéa, est perçu avec une précision nettement supérieure aux directions périphériques. Cette coïncidence des zones de meilleure résolution spatiale entre les modalités auditive et visuelle est à souligner. En effet il est généralement admis que, dans le cadre de la perception de l'espace, la complémentarité des modalités auditive et visuelle est basée sur leur fonctionnement dans des domaines disjoints : un stimulus visuo-auditif peut être localisé en dehors du champ de vision par l'audition puis, une fois l'attention visuelle portée sur cet objet, analysé par le système visuel.

Imperfections du système auditif Ce chapitre a permis de mettre en évidence un certain nombre de distorsions que le système auditif opère sur le signal acoustique. Citons par exemple les résonances et délais temporels induit par le pavillon ou les phénomènes de masquage introduits par la cochlée. La psychoacoustique met également en évidence une relation non-linéaire entre certaines quantités physiques et la manière dont nous les percevons : le flou de localisation pour la perception spatiale,

l'échelle de Bark utilisée pour la perception de l'intensité sonore, ou encore l'échelle de Mel pour celle de la fréquence fondamentale. De plus ces distorsions, notamment celles produites par la tête et les pavillons sont d'une importance fondamentale pour parvenir à localiser une source sonore.

Cette observation est à mettre en parallèle avec les imperfections du système visuel, dans lesquelles Poincaré (1895), et O'Regan (2011) à sa suite, trouvent un argument pour la théorie sensorimotrice. Ainsi l'image projetée sur la rétine est inversée (dans le sens haut-bas) et présente une déformation sphérique que nous ne percevons pas. La zone aveugle au milieu de la rétine, à l'endroit où convergent les fibres nerveuses, ne nous est également pas perceptible, de même que les différents canaux vasculaires irriguant l'oeil, le flou visuel engendré par les saccades oculaires nous est enfin complètement invisible. Cette idée que nos systèmes perceptifs seraient dénués d'imperfections et nous offriraient une reproduction fidèle de notre environnement est en fait issue de la vision traditionnelle et *bottom-up* de la perception. Ainsi ce constat mène à reconsidérer la perception non pas du point de vue de la qualité de reproduction qu'elle peut offrir, mais par les capacités d'interactions qu'elle implique : percevoir n'est pas se représenter une scène, c'est interagir avec elle.

Chapitre 3

Robotique binaurale pour la localisation de sources sonores

Sommaire

La localisation de sources sonores est d'une importance majeure dans les capacités auditives des animaux comme l'a montré le chapitre 2. C'est également l'application principale proposée par les modèles auditifs existant en robotique, les deux autres fonctions, complémentaires de cette première, étant la séparation et la reconnaissance de sources sonores. Ce chapitre présente ainsi un état de l'art des modèles auditifs binauraux appliqués à la localisation de sources sonores en robotique. Nous commencerons par exposer quelques propriétés communes aux traitements auditifs,

d'abord de manière générale puis spécifique à la robotique. Nous proposerons ensuite un état de l'art portant sur les solutions auditives binaurales proposées dans la littérature. Cet état de l'art se focalisera sur les « plus-values » apportées par le contexte robotique, que nous décomposons en deux sous-ensembles : la perception active, intégrant le mouvement au coeur du processus de perception, et l'apprentissage construit à partir de l'expérience du robot. Nous présenterons finalement quelques applications en relation étroite avec la localisation concernant la séparation de sources, leur reconnaissance ou encore l'intégration audiovisuelle. La discussion concluant ce chapitre sera l'occasion de faire apparaître un clivage entre les méthodes robotiques essentiellement *bottom-up*, et l'organisation biologique du système auditif fondamentalement *top-down*.

3.1 Systèmes auditifs en robotique

Après la présentation d'un cadre général à l'élaboration de systèmes auditifs artificiels, ce paragraphe revient plus spécifiquement à l'audition robotique en présentant les deux principales approches introduites dans la littérature et les contraintes spécifiques au contexte robotique auxquelles ces méthodes doivent faire face.

3.1.1 Modélisation du système auditif

Qu'il soit dédié à la robotique ou à toute autre fin applicative, un modèle de système auditif se compose de la manière la plus générale de 4 étapes successives de traitements opérant sur les données fournies par un ou plusieurs microphones (Lyon, 2010). Un étage de *prétraitement* rempli tout d'abord le rôle du système auditif périphérique : il filtre le signal temporel et en extrait les informations fréquentielles. La seconde étape du traitement consiste en l'*extraction* des indices de base utilisés par le système, comme les indices d'ITD ou d'ILD dans le cas de la localisation. L'étape suivante a pour fonction de fournir une *interprétation* de ces données brutes, fournissant ainsi des informations sur le signal source ou sur le contexte environnemental, comme la position d'une source dans le cas de la localisation. Nous pouvons enfin considérer une dernière phase de traitement, au plus haut niveau d'abstraction, consistant en un module de *décision* et d'*intégration*. Sa fonction peut être, dans un cadre robotique, la gestion d'une réponse comportementale ou l'intégration avec d'autres modalités sensorielles.

3.1.2 Approches et applications

Deux paradigmes principaux sont proposés dans la littérature pour l'audition en robotique : les solutions à base d'antennes de microphones et les systèmes binauraux. Une antenne, tout d'abord, est composée de plusieurs microphones qui peuvent être organisés selon diverses géométries (en ligne, en cercle, etc). La redondance d'information apportée par le nombre des capteurs permet d'accroître la robustesse de l'analyse auditive (Van Trees, 2002) et ce paradigme est employé dans de nombreuses applications, notamment la localisation de sources (Julian *et al.* , 2004; Brutti *et al.* , 2008), mais également l'estimation du nombre de sources actives ou encore l'adaptation d'algorithmes de traitement de la parole au contexte robotique (Argentieri *et al.* , 2013). En termes de localisation de sources sonores, la contribution principale de ce domaine de recherche est l'extension à la robotique de méthodes classiques de

localisation à base d'antennes, basées notamment sur des approches haute résolution (Argentieri & Danès, 2007) ou sur le *beamforming* (Nakajima *et al.* , 2009).

Le deuxième paradigme en robotique est celui de l'audition binaurale (Argentieri *et al.* , 2013). L'organisation binaurale du système auditif est celle qui a émergé de l'évolution et a à ce titre fait l'objet de très nombreuses modélisations en dehors du contexte robotique (Fay & Popper, 2010; Blauert, 2013). Les différents modèles binauraux proposés dans la littérature peuvent généralement se distinguer selon leur degré de biomimétisme. D'une part les modèles computationnels les plus fins sont en général utilisés pour valider ou infirmer des hypothèses sur les structures biologiques qu'ils modélisent. D'autre part des modèles plus synthétiques, qu'ils soient inspirés de structures biologiques ou de méthodes de traitement du signal, offrent de nombreuses applications. Nous pouvons citer parmi celles-ci la génération d'environnements acoustiques virtuels, la conception de prothèses auditives traitant la surdité ou les pertes auditives, la reconnaissance de la parole ou du locuteur et bien sur l'audition robotique (Blauert, 2013).

3.1.3 Contraintes propres à la robotique

Comme nous venons de le voir, les approches de modélisations du système auditif sont variées. Néanmoins, considérant le contexte de la robotique, tous les modèles ont à prendre en compte un ensemble de contraintes propre au domaine (Argentieri *et al.* , 2013) parmi lesquelles nous pouvons citer :

1. La contrainte embarquée :

   Une solution auditive, et en particulier la couche capteur, doit pouvoir être embarquée sur une plateforme robotique. L'économie de ressources proposée par les solutions binaurales présente ici un intérêt évident, à l'inverse des antennes de microphones dont l'efficacité croît en fonction du nombre et de l'espacement des capteurs qui la composent.

2. La contrainte temps-réel :

   Une application robotique, et *a fortiori* une application dédiée à la perception, doit opérer en temps réel pour proposer des résultats utilisables (Pressnitzer & Gnansia, 2005). Cependant, selon l'algorithme utilisé, certaines approches nécessitent des calculs très lourds. Ceci est encore une fois particulièrement vrai pour les solutions à base d'antenne de microphones où de coûteuses corrélations croisées sont estimées entre des canaux multiples. Cette contrainte d'un modèle computationnellement efficace fera l'objet d'une attention constante lorsque nous présenterons notre modèle binaural au chapitre 4.

3. La contrainte morphologique :

   Les capacités motrices d'un robot permettent de mettre en place des stratégies de perception active qui, comme nous l'avons développé au paragraphe 2.1.3, sont exploitées par notre propre système auditif. La morphologie de l'oreille externe est de plus un élément à part entière du système auditif qui joue un rôle fondamental pour la localisation de sources sonores, notamment grâce à l'ombre acoustique causé par la tête et au filtrage opéré par les pavillons.

4. La contrainte environnementale :

   L'environnement dans lequel évolue un robot est, sauf cas particuliers, largement dynamique et imprévisible. Bruits de fonds, bruits impulsionnels, réverbérations de l'environnement et non-stationnarité des sources sont autant

d'éléments perturbateurs auxquels le système robotique doit être robuste. De plus le robot lui-même, de par l'action de ses moteurs, peut émettre des sons susceptibles de corrompre le bon fonctionnement du système. Ces ego-bruits, ou bruits auto-générés, s'ils ne constituent pas un problème fondamental sont en revanche une difficulté pratique qui doit être pensée dès la conception du système (voir à ce sujet le paragraphe 3.2.1.2).

Alors que les deux premières contraintes constituent des problèmes purement techniques et technologiques, les deux dernières revêtent un caractère beaucoup plus fondamental et impliquent la conception de nouveaux formalismes et modèles. Malgré les quelques avantages que propose l'approche binaurale dans la réponse à ces contraintes, une implémentation robotique à base d'antenne de microphones et tout à fait possible. Néanmoins cette seconde approche s'avère plus complexe à mettre en oeuvre et plus éloignée de la biologie que la solution binaurale. Le propos central de cette thèse est de plus d'explorer la théorie sensorimotrice appliquée à la localisation active de sources sonores, ce qui justifie d'autant plus le choix d'un modèle auditif relativement proche de la biologie et donc notre focalisation sur les modèles binauraux.

3.2 Robotique binaurale pour la localisation

Ce paragraphe propose un état de l'art consacré à la robotique binaurale appliquée à la localisation de sources sonores. Les modèles auditifs robotiques se distinguent par leur prise en compte des contraintes évoquées précédemment mais également par le fait qu'ils peuvent explorer des stratégies de perception actives propres au contexte robotique, de même qu'ils peuvent tirer profit de l'expérience emmagasinée au travers de méthodes d'apprentissage artificiel. Il est difficile de proposer une synthèse organisée des différentes méthodes existantes et nous avons choisi de distinguer 4 catégories : l'oreille externe, l'estimation des indices binauraux, les méthodes actives pouvant être de plus ou moins haut niveau, et enfin les méthodes à base d'apprentissage regroupant les processus de plus haut niveau. Ainsi cet état de l'art ne tient donc pas à répertorier les différents traitements binauraux existants mais plutôt à comprendre de quelle manière ils peuvent être exploités avec profit dans un contexte robotique. Quelques recherches menées sur des capteurs binauraux non robotisés ou à partir de simulations numériques sont également présentées, dans la mesure ou elles contribuent de manière pertinente au domaine de l'audition robotique.

3.2.1 Oreille externe

Ce paragraphe présente les différents modèles d'oreilles externes développés pour la robotique. Le capteur binaural ne se limite en effet pas à une simple paire de microphones puisque l'oreille externe, qui fournit des indices acoustiques essentiels à la localisation, est un élément à part entière d'un système binaural. Quelques architectures de pavillons sont d'abord décrites puis des méthodes de réduction des bruits auto-générés sont présentées.

3.2.1.1 Pavillons artificiels

Quelques pavillons artificiels ont été proposés pour la localisation en robotique, trois exemples significatifs étant illustrés Fig. 3.1. En s'inspirant des caractéristiques

(a) Shimoda *et al.* (2007)

(b) Kumon & Noda (2011)

(c) Deleforge & Horaud (2011)

Figure 3.1 – Exemples de pavillons artificiels utilisés en robotique. (a) ICub équipé de pavillons simples (Shimoda *et al.* , 2007). (b) Oreille externe active, souple et déformable (Kumon & Noda, 2011). (c) Mannequin binaural monté sur une unité pan-tilt également équipé d'un système de vision binoculaire (Deleforge & Horaud, 2011).

biologiques de l'oreille externe, l'objectif est de produire des HRTF à partir desquelles sont extraits puis analysés les indices spectraux ainsi créés. Hörnstein *et al.* (2006), Shimoda *et al.* (2007), Rodemann *et al.* (2008) ou encore Finger *et al.* (2010) se basent par exemple sur les indices spectraux créés par des pavillons aux formes simples pour proposer une méthode de localisation en azimut et en élévation. On peut également citer les pavillons artificiels présents sur des robots plus complets, tel l'humanoïde iCub (Park & Hwang, 2007) ou le robot Asimo (Yamamoto *et al.* , 2006). Lee *et al.* (2008) exploitent une asymétrie entre les deux oreilles pour fournir, en plus de la localisation en élévation, une méthode permettant de lever l'ambiguïté avant-arrière. Dans le cadre de la modélisation de l'écholocation chez la chauve-souris, une série d'études (Walker *et al.* , 1998; Peremans & Muller, 2000; Peremans & Reijniers, 2005) a proposé un modèle de pavillon de chauve souris dont la morphologie et le contrôle actif permettent de générer des indices spectraux spécifiques à l'écholocation. Plus en prospective, Rodemann (2011) proposent également une méthode d'extraction des indices spectraux à partir d'un forme de pavillon arbitraire tandis que Kumon & Noda (2011) proposent une oreille externe souple et déformable et établissent un lien entre mobilité du pavillon et indices spectraux.

3.2.1.2 Réduction des bruits auto-générés

Les bruits causés par les moteurs d'un robot en mouvement constituent un problème important dans les applications liées à l'audition artificielle, qu'elles soient consacrées à la localisation ou à la reconnaissance de la parole (Nishimura *et al.* , 2006; Ince *et al.* , 2010). De par la proximité des moteurs avec les microphones, le

bruit généré peut en effet être capté avec une intensité importante. Certaines applications doivent de plus compter avec des bruits indépendants des mouvements du robot, tels que ceux d'origine électrique générés par des transformateurs ou encore ceux causés par les ventilateurs. Plusieurs solutions ont été proposées dans le but d'atténuer ces différents bruits d'origine interne. La solution la plus simple, retenue par exemple par Murray *et al.* (2004), est d'adopter une stratégie dite du *stop and hear* qui consiste à désactiver l'écoute durant le mouvement du robot. Cette approche atteint toutefois ses limites dès lors que la modalité auditive requière d'être opérationnelle pendant le déplacement du robot.

Ainsi Nakadai *et al.* (2000b,c) proposent un système binaural dont chaque oreille est équipée d'une paire de microphones. Le premier microphone, situé à l'extérieur de la coque, est dédié à la perception des sons environnementaux tandis qu'un second microphone, positionné à l'intérieur de la structure du robot, capte principalement les bruits générés par les moteurs de la tête. Ce microphone permet de détecter les sous-bandes spectrales du signal dans lesquelles le bruit interne est dominant, ces sous-bandes sont ensuite simplement ignorées par les phases de traitement en aval. La suppression pure et simple de ces sous-bandes spectrales est bien trop réductrice pour nombre d'applications si bien qu'une approche plus complexe, basée sur la soustraction spectrale, a également été mise en oeuvre (Ito *et al.* , 2005; Nishimura *et al.* , 2006; Ince *et al.* , 2010, 2011a,b). Une base de donnée contenant les spectres de bruits générés par différents mouvements est ainsi utilisée. Avec la connaissance du mouvement opéré par le robot, le bon *template* est sélectionné dans la base puis soustrait au signal enregistré. Cette méthode est relativement lourde à implémenter puisqu'elle requiert à la fois la construction d'une large base de données et une phase d'apprentissage hors-ligne. Pour faire face à ces limitations Ince *et al.* (2011c) proposent une méthode d'apprentissage en ligne de ces *templates* qui permet également une adaptation de la base de donnée aux conditions environnementales.

3.2.2 Extraction des indices binauraux

Ce paragraphe décrit le principe de l'estimation de l'ITD et de l'ILD généralement retenue par les modèles binauraux, ce principe n'étant pas réservé à la robotique. Bien que ce principe général soit souvent le même, il existe une très forte variabilité des méthodes dans la littérature, avec notamment l'ajout de pré ou post-traitements différents selon les auteurs et les applications visées.

3.2.2.1 Différence interaurale de temps

L'ITD, qui mesure le retard interaural entre les signaux provenant des oreilles gauche et droite, peut être estimé mathématiquement comme le décalage temporel maximisant la corrélation croisée des signaux en entrée (Knapp & Carter, 1976). Considérant ainsi les signaux temporels $l(t)$ et $r(t)$ correspondant aux signaux gauche et droit respectivement, leur corrélation croisée $\Gamma_{l,r}(\tau, t)$ sur une fenêtre T s'exprime basiquement comme :

$$\Gamma_{l,r}(\tau, t) = \int_{t-T}^{t} l(u)r(u-\tau)du, \tag{3.1}$$

où τ est le délai pour lequel est estimé la corrélation. Une fois la corrélation calculée pour différents délais, le retard interaural $\tau_{itd}(t)$ correspondant à l'ITD au temps t

est alors :

$$\tau_{itd}(t) = \underset{\tau}{\text{argmax}}(\Gamma_{l,r}(\tau, t)). \tag{3.2}$$

Nous l'avons déjà évoqué, cette différence de temps d'arrivée τ_{itd} peut se voir, par simple raisonnement géométrique, comme fonction de l'azimut θ de la source considérée. Considérant en première approximation la tête comme acoustiquement transparente (ne provoquant aucune atténuation ou réflexion), nous avons ainsi :

$$\tau_{itd} = \frac{d}{c}\sin\,\theta, \tag{3.3}$$

où d est la distance interaurale et c la vitesse du son. Si l'hypothèse de la neutralité de la tête s'avère bonne pour les basses fréquences, les perturbations deviennent trop importantes au-delà de 500 Hz (Stern *et al.* , 2006) et une deuxième approximation, prenant en compte la présence de la tête, devient nécessaire. Considérant maintenant que l'onde sonore voyage le long de la tête, assimilée à une sphère, avant de parvenir aux tympans, le délai τ_{itd} s'exprime ainsi :

$$\tau_{itd} = \frac{d}{2c}(\theta + \sin\,\theta). \tag{3.4}$$

Considérant ainsi la vitesse du son $c = 340$ m.s^{-1} et une tête de distance interaurale $d = 0.18$ m, ce qui correspond à la valeur moyenne chez l'homme (Algazi *et al.* , 2002), le délai maximal atteint pour une source à 90° (c'est-à-dire ici $\theta = \pi/2$) est égal à $\tau_{max} = 530$ μs par l'Eq. 3.3 et à $\tau_{max} = 680$ μs par l'Eq. 3.4.

3.2.2.2 Différence interaurale d'intensité

L'extraction de l'ILD retient généralement moins l'attention que l'extraction de l'ITD dans les modèles de localisation binaurale. Les modèles exploitant uniquement l'ILD, comme celui de Smith (1992) ou notre modèle de localisation active décrit au paragraphe 6 (voir également (Bernard *et al.* , 2010a)), sont exceptions dans la littérature. Ceci est vraisemblablement dû au fait qu'il est difficile de remonter à l'azimut d'une source à partir de la différence d'intensité seulement. Cette opération nécessite en effet des *a priori* sur la source et/ou sur l'agent. L'ILD est ainsi généralement utilisé conjointement à l'ITD, par exemple par Faller & Merimaa (2004), Liu *et al.* (2008) ou encore Raspaud *et al.* (2010). Bien qu'il existe des modèles de LSO estimant l'ILD de manière bioinspirée, tel celui de Calmes (2009), proposant un modèle neurocomputationnel du LSO de la chouette effraie, l'ILD s'estime classiquement comme le ratio des intensités gauche et droite, soit directement sur le signal temporel soit, pour les modèles à base de transformée de Fourier, à partir de sa représentation fréquentielle. Nous venons de dire que l'ILD se calcule comme le ratio des intensités gauche et droite. De manière équivalente, ce ratio peut également s'exprimer comme une différence de logarithmes, expliquant que le terme ILD indique une *différence* d'intensité. Considérant ainsi les signaux gauche et droit $l(t)$ et $r(t)$, et leur transformées de Fourier respectives $L(f)$ et $R(f)$, l'ILD τ_{ild} est donné en fonction de la fréquence f comme :

$$\tau_{ild}(f) = 20\log_{10}(\frac{|L(f)|}{|R(f)|}). \tag{3.5}$$

3.2.3 Perception active

Le chapitre précédent nous a montré la variété des comportements actifs liés à la localisation de sources sonores par les animaux. En robotique nous identifions trois grandes approches. Premièrement les modèles passifs au-dessus desquels est rajoutée une dimension active constituent l'approche la plus simple. Les méthodes multiposes, ensuite, permettent d'exploiter les mouvements de l'axe interaural pour déterminer la position d'une source de manière analytique. Enfin, les modèles de comportements réflexes proposent de reproduire des capacités de perception active retrouvées dans la nature grâce à un couplage sensorimoteur bas-niveau. Nous verrons également dans le paragraphe 3.2.4 que des méthodes à base d'apprentissage peuvent également exploiter des stratégies actives, notamment pour le suivi de sources en mouvement.

3.2.3.1 Adaptation de modèles passifs

Certains modèles binauraux issus de la modélisation dans un contexte passif ont été implémentés en robotique. Néanmoins ces modèles ne tirent pas nativement profit des caractéristiques apportées par les mouvements et le système moteur du robot. Ainsi Murray *et al.* (2004) portent en robotique un modèle binaural très classique, où une corrélation croisée est opérée directement sur le signal temporel. Les performances de localisation atteignent une résolution d'environ 1.5° en azimut. Pour ajouter une composante active à ce modèle les auteurs proposent un module moteur qui permet au robot de s'orienter vers l'azimut estimé par corrélation. Le modèle résout l'ambiguïté avant-arrière en mesurant le gradient d'intensité entre deux mesures successives de l'ILD, chacune séparée d'une rotation de quelques degrés. Néanmoins ce modèle n'implémente à proprement parler aucun couplage sensorimoteur permettant de le qualifier de modèle actif, puisque l'action effectuée *a posteriori* est ici indépendante du signal perçu.

Trifa *et al.* (2007) comparent 4 méthodes d'extraction d'ITD dans le cadre d'une localisation de signaux de parole, d'abord en simulation puis lors d'une expérimentation robotique. Les méthodes comparées sont, sur la base d'un filtrage gammatone (voir paragraphe 4.1.2.2), la corrélation croisée généralisée (GCC), et sur la base d'une transformée de Fourier (FFT), la GCC, la GCC avec transformée de phase (GCC-PHAT), et le critère de délai de Moddemeijer (MODD). Le calcul de la GCC est présenté par l'Eq. 3.2, la GCC-PHAT y ajoute un prétraitement de la phase permettant d'augmenter la robustesse aux réverbérations (Knapp & Carter, 1976) tandis que MODD est une approche basée sur la théorie de l'information cherchant à maximiser la probabilité de la position de la source à partir d'une mesure de l'information mutuelle entre les signaux gauche et droit (Moddemeijer, 1988; Trifa *et al.* , 2007). Le modèle à base de gammatone est plus efficace pour des SNR supérieurs à 1 dB et des sources à large spectre. Lorsque le spectre est plus étroit les méthodes basées FFT l'emportent, GCC-PHAT étant la plus précise et MODD la plus stable. Calmes *et al.* (2007) fournissent également une étude poussée sur l'utilisation de l'ITD pour la localisation en azimut sur une plateforme robotique binaurale. L'ITD est calculé par un modèle dérivé de celui de Jeffress (1948), un modèle bioinspiré également utilisé dans cette thèse et qui sera décrit en détail au paragraphe 4.2.1. L'efficacité de l'algorithme est démontrée pour différents types de signaux, mais les performances chutent en présence de tons purs ou de réverbération importante. Ainsi la localisation d'un bruit blanc se dégrade à partir d'un SNR de -20 dB en conditions anéchoïdes et à partir d'un SNR de -10 dB en conditions réverbérantes.

Rodemann *et al.* (2006) et Heckmann *et al.* (2006) proposent un modèle de localisation bioinspiré robuste aux environnements bruités et réverbérants. Ce modèle, reproduisant l'effet de précédence, est relativement complet puisqu'il intègre les 3 indices binauraux que sont la différence de temps, d'intensité et d'enveloppe. L'utilisation de 3 indices différents est motivée par la recherche de robustesse. Les indices sont calculés par corrélation à partir d'une représentation des passages par zéro du signal d'entrée. Une performance temps-réel est atteinte par une exécution parallèle sur plusieurs machines et permet aux auteurs l'expérimentation du modèle sur une tête humanoïde. Le modèle binaural commande ainsi l'asservissement de la tête, ce qui permet au robot de s'orienter vers une source sonore en fonction de sa position estimée et de la suivre si elle est en mouvement. Ce modèle s'avère robuste aux conditions échoïques mais reste sensible au bruit. Une expérimentation sur la perception de la distance a également été effectuée à partir de ce modèle (Rodemann, 2010). Les auteurs montrent ainsi la pertinence des indices basés sur l'intensité pour cette tâche, à l'inverse des indices purement spectraux. Enfin Finger *et al.* (2010) proposent une méthode de localisation en azimut et en élévation qui inclut pavillons artificiels, décomposition FFT et estimation de l'ITD par GCC et de l'ILD, la fusion des deux indices étant effectuée par moindre carré. Les performances obtenues par ce modèle dans une tâche de localisation atteignent 2.5° en azimut et 11° en élévation.

3.2.3.2 Méthodes multiposes

Dans un travail qui semble être le premier du genre, Reid & Milios (2003), qui nomment d'ailleurs leur approche "audition active", proposent un modèle de localisation binaurale qui exploite les rotations en azimut et en élévation d'une plateforme à 2 DOF sur laquelle est montée une paire de microphones omnidirectionnels (Fig. 3.2). L'ITD est calculé par corrélation croisée sur des signaux provenant de deux poses différentes, permettant ainsi d'estimer l'intersection des deux cônes de confusion générés. Les auteurs proposent également une méthode de localisation basée sur l'intensité du signal, qui est ainsi mesurée à différentes positions (2 ou 3 suffisent en pratique) puis corrélée à la réponse directionnelle des microphones par minimisation quadratique. Cette méthode implique néanmoins la connaissance *a priori* de cette réponse directionnelle et, de part l'approche multipose retenue et la nécessité de plusieurs mesures successives, repose sur l'hypothèse d'une source unique, stationnaire et immobile.

Kneip & Baumann (2008) ont par la suite étendu la méthode de Reid & Milios (2003) basée sur l'intersection des cônes de confusion. Ils ont en effet proposé une formalisation du problème de la localisation binaurale exploitant rotations et translations de l'axe interaural. Les auteurs démontrent ainsi qu'une unique rotation, et donc deux mesures seulement, est suffisante pour déterminer l'azimut et l'élévation d'une source sonore, à une confusion près, qu'elle soit avant-arrière ou haut-bas selon la rotation effectuée. De plus l'ajout des translations permet de résoudre théoriquement complètement le problème de la localisation, y compris en distance, grâce au calcul de la parallaxe de mouvement. Cette méthode permet donc d'estimer les coordonnées cartésiennes de la source par rapport au robot. Une expérience menée sur une plateforme robotique permet d'atteindre une résolution moyenne de 10° en azimut et en élévation, et de 0.5 m en distance. Bien que cette méthode présente des résultats encourageants et une base théorique explicite, elle est fortement limitée par la contrainte d'une source sonore stationnaire et immobile, ce qui réduit son intérêt pour une utilisation dans un environnement complexe.

FIGURE 3.2 – Unité pan-tilt équipée de deux microphones omnidirectionnels exploitant la méthode de localisation multiposes, pionnière en audition active (Reid & Milios, 2003).

3.2.3.3 Comportements réflexes

A l'inverse de la méthode analytique précédente, les comportements réflexes mettent directement en relation la perception et l'action dans une boucle sensorimotrice simple. Les comportements réflexes présentent ainsi une alternative aux solutions plus classiques pour la localisation. En effet il n'y a pas de localisation explicite de la source sonore, c'est-à-dire qu'aucun angle ou direction n'est estimé *a priori*. La source est à l'inverse localisée *a posteriori*, c'est-à-dire une fois qu'un mouvement a été effectué. Ces méthodes sont toutefois limitées par leur caractère principalement réactif et, même après une phase d'apprentissage présente dans certains modèles, la localisation est impossible sans mouvement.

Dans ce qui constitue un travail précurseur dans le domaine de la perception active, Braitenberg (1986) a proposé une série de modèles très simples dans lesquelles une paire de capteurs visuels est directement reliée à une paire de roues . Selon la relation inhibitrice ou excitatrice des connexions et leur branchement ipsi ou contralatéral, le comportement résultant peut devenir, pour un même stimulus, un comportement d'approche ou d'éloignement, comme illustré Fig. 3.3. Le comportement de ces véhicules a également été formalisé par Rañó (2007, 2012).

Les comportements réflexes sont largement présents chez les animaux, comme l'a montré le paragraphe 2.1.3. Ils ont également fait l'objet de modélisations dans le cadre d'une approche biomimétique. Une série d'études (Rucci & Wray, 1999; Rucci *et al.* , 1999, 2000) a ainsi proposé un modèle computationnel du comportement d'orientation de la chouette effraie avec un degré de biomimétisme élevé et une modélisation des différentes structures cérébrales impliquées. Ce modèle, implémenté sur une plateforme robotique, permet l'apprentissage de ce comportement d'orientation sur la base d'une supervision visuelle. Après apprentissage la précision du mouvement d'orientation atteint 1° environ. Le modèle est de plus capable de s'adapter à des altérations de ses entrées sensorielles ou de ses sorties motrices.

Suivant la même démarche biomimétique, une autre série d'études s'est penchée sur la modélisation du comportement de phonotaxie du grillon (Webb & Scutt, 2000; Reeve & Webb, 2003; Horchler *et al.* , 2004; Reeve *et al.* , 2005). Ce comportement,

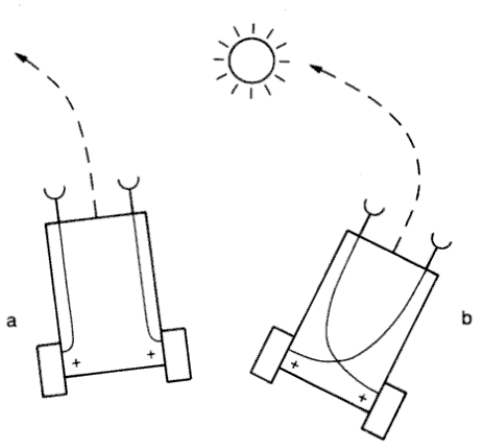


Figure 3.3 – Véhicules de Braitenberg (1986) composés de deux senseurs reliés à deux actionneurs. (a) Deux connexions excitatrices dans une configuration ipsilatérale engendrent un comportement d'éloignement de la source. (b) La configuration excitatrice contralatérale provoque à l'inverse un comportement d'approche. D'après Braitenberg (1986)

permettant au grillon femelle de se diriger vers son partenaire, est spécialisé dans la reconnaissance d'un chant spécifique à l'espèce. Le système moteur est contrôlé directement par des indices liés à l'intensité du signal. Ce modèle à base de réseaux de neurones a été implémenté sur une plateforme robotique et expérimenté avec succès dans le milieu naturel du grillon. Toujours concernant la phonotaxie du grillon Damper & French (2003) s'intéressent plus particulièrement à l'évolution artificielle du réseau de neurones associant les entrées sensorielles aux sorties motrices. À partir d'un système binaural Manoonpong *et al.* (2005) proposent également un réseau de neurones optimisé par un algorithme évolutionniste. Ce réseau est dédié au contrôle sensorimoteur de la marche sur un robot quadrupède et permet au robot de se diriger vers l'origine d'une source sonore.

3.2.4 Apprentissage

Les méthodes d'apprentissage sont fréquemment utilisées dans les applications dédiées à la localisation de sources sonores car elles permettent de mettre en relation les valeurs des indices binauraux avec une commande motrice, une position dans l'espace, ou encore pour l'apprentissage des HRTF. L'apprentissage peut également être utilisé pour la mise en place de stratégies actives plus complexes que celle vues au paragraphe précédent, en cherchant par exemple à prendre en compte la dynamique d'une source et l'évolution de sa position au fil du temps.

3.2.4.1 Probabilités *a posteriori*

Une approche probabiliste peut ainsi être suivie dans le but d'accroître la robustesse d'un système aux perturbations de l'environnement. Pinho *et al.* (2008) proposent ainsi un modèle bayesien pour la localisation en azimut, en élévation et en distance. Basé sur un système périphérique à base de gammatone (Patterson *et al.* , 1995) et sur l'extraction des indices d'ITD et d'ILD basés sur la cohérence interaurale (Faller & Merimaa, 2004), ce modèle consiste en l'apprentissage d'une représentation

de l'espace auditif appelée carte volumétrique bayesienne. Deleforge & Horaud (2012) proposent également un modèle probabiliste mélangeant localisation et séparation de sources dans une approche inspirée de l'algorithme d'espérance-maximisation (EM). La procédure proposée alterne ainsi entre une étape de séparation (espérance), assignant une source à chaque point temps-fréquence du spectrogramme par estimation des probabilités *a posteriori*, et une étape de localisation (maximisation) qui permet de maximiser ces probabilités et d'en déduire la position des sources considérées.

En se basant sur l'apprentissage supervisé d'un modèle de mélanges gaussiens (GMM), May *et al.* (2011) s'intéressent à la localisation dans un contexte réverbérant et multisource. Les paramètres du GMM sont appris par EM à partir de vecteurs d'ILD et d'ITD obtenus à partir d'un filtrage à base de gammatone ou de FFT, l'ITD étant estimé par GCC. Les sources sonores sont des signaux de parole à différentes distances et azimuts. Les auteurs fournissent une étude poussée investiguant l'influence de différents paramètres sur les performances de localisation. Ils démontrent ainsi que ce modèle offre de bonnes performances en présence de configurations source/auditeur absentes de l'ensemble d'apprentissage, prouvant ainsi une certaine capacité de ce modèle à la généralisation. Ce modèle est de plus capable d'estimer le nombre de sources actives. L'approche probabiliste permet de plus de prendre en compte l'incertitude des indices binauraux causée par la réverbération, le contexte multisource ou les changements de configuration. À ce titre, les auteurs remarquent que la distance entre la source et l'auditeur est un paramètre sensible en présence de réverbération, l'écart de performance entre des distances à 1.5 m et 2 m étant significatif.

3.2.4.2 Filtrage de Kalman

A l'inverse des méthodes faisant l'hypothèse d'une source sonore immobile, certains auteurs proposent d'intégrer les mouvements relatifs de la source dans le processus de localisation. Ainsi Kumon & Uozumi (2011) proposent une méthode à base de filtrage de Kalman pour la localisation d'une source en mouvement et l'illustrent par la simulation d'une plateforme robotique binaurale. Le filtrage permet l'estimation d'une commande motrice optimale permettant, au fil des itérations, de minimiser la matrice de covariance associée au filtre et donc de minimiser l'erreur de localisation.

Le modèle proposé par Portello *et al.* (2011), également basé sur un filtrage de Kalman, cherche à intégrer les changements acoustiques induits par le mouvement de la source et/ou du robot afin de mettre à jour l'estimation de la position de la source. L'exploitation des mouvements relatifs permet avec cette méthode de lever l'ambiguïté avant-arrière inhérente à l'utilisation de l'ITD. Le filtrage permet une estimation précise du mouvement après 2 secondes environ et entraîne une localisation très stable en azimut, avec une variance d'erreur de localisation de 3°. L'estimation en distance est moins précise, avec une variance de 1 m environ.

3.2.4.3 Modèles connexionnistes

Les modèles connexionnistes peuvent être utilisés dans un contexte passif, par exemple pour l'estimation de la direction (Backman & Karjalainen, 1993), l'estimation de l'ITD (Glackin *et al.* , 2010) ou l'apprentissage des HRTF (Goodman & Brette, 2010b,a). Ces modèles sont également appliqués à la robotique et, sur la base du modèle proposé par Rodemann *et al.* (2006) et décrit précédemment, Rodemann *et al.* (2007) proposent ainsi une méthode d'apprentissage en ligne d'une carte au-

ditorimotrice à partir d'un réseau de neurones. Cette carte consiste en un mapping linéaire qui met en relation la valeur moyenne de l'ITD avec l'azimut de la source. Deux mesures d'ITD sont nécessaires, le mouvement entre ces deux mesures pouvant être quelconque. Cette méthode repose sur l'hypothèse d'une source immobile. La méthode est testée sur une tête binaurale et le modèle atteint de bonnes performances après un apprentissage de 400 itérations, soit 2 heures environ. À noter que des performances similaires sont obtenues avec l'ILD.

Berglund & Sitte (2005) et Berglund *et al.* (2008) proposent également un modèle binaural pour la localisation en 3 dimensions basé sur une carte auto-organisatrice (SOM). 4 indices binauraux sont extraits d'une décomposition FFT, l'ITD, l'IPD, l'ILD et sa version normalisée la RILD. Après une phase d'apprentissage dont la lenteur est un point négatif selon les auteurs, la SOM permet de représenter les indices binauraux en entrée en basse dimension. Un mécanisme d'apprentissage par renforcement permet ensuite d'exprimer ces représentations en termes de commandes motrices. Cette méthode utilise de plus les mouvements à court terme de la source pour parvenir à une localisation en 3 dimensions. Une expérimentation robotique est présentée sur le robot Aibo. La SOM est entraînée à partir de 10000 échantillons, les sources sonores étant disposées à divers azimuts et distances. L'apprentissage par renforcement nécessite quant à lui 2000 itérations pour converger vers des résultats convenables. De part cet apprentissage d'une représentation en basse dimension de l'espace des indices, cette étude se rapproche des modèles se basant sur la théorie sensorimotrice qui seront décrits à partir du chapitre 5.

3.3 Applications basées sur la localisation

Les méthodes décrites précédemment ont pour finalité principale la localisation de sources. Il existe cependant des applications allant au-delà de la simple localisation et permettant par exemple la séparation de sources sur la base de critères spatiaux ou encore l'intégration de la modalité auditive avec la vision, permettant la localisation de sources audiovisuelles, ou *a minima* l'expression de l'espace auditif en termes de coordonnées dans l'espace visuel. Ce paragraphe présente ainsi des applications robotiques dont la base est un modèle de localisation binaural et dont la finalité applicative est liée à l'intégration audiovisuelle de sources tout d'abord, et à la reconnaissance et la séparation de sources ensuite.

3.3.1 Intégration audiovisuelle

Nous avons vu au paragraphe 2.3.1 que le système nerveux procède à une fusion des indices audiovisuels au niveau de SC et que, à ce niveau, l'espace auditif est recalé sur l'espace visuel, permettant ainsi de localiser un stimulus dans l'espace de manière amodale. Schauer & Gross (2003) et Nguyen *et al.* (2010) proposent ainsi un modèle de fusion audiovisuelle inspiré par SC dans laquelle est construite une carte bimodale. Une propriété intéressante observée dans SC que les auteurs modélisent est l'amélioration de la réponse bimodale par rapport à une réponse unimodale considérée isolément. Youssef *et al.* (2011, 2012a) proposent également une méthode de localisation dans laquelle l'apprentissage de la position d'une source est effectué grâce à une supervision visuelle à partir de l'ITD et de l'ILD. La source est ainsi localisée dans le champ visuel en coordonnées pixels.

Cette fusion des espaces auditif et visuel permet de considérer des stimuli bimodaux. Avec pour objectif de proposer un système de « dialogue » homme-robot, Nakadai *et al.* (2000a,c) utilisent ainsi la géométrie épipolaire pour la localisation auditive et visuelle dans un même formalisme, la géométrie épipolaire ayant été initialement introduite en vision artificielle pour la reconstruction de scènes tridimensionnelles à partir de points de vues 2D. La perception active est également utilisée par ce modèle auditif pour l'orientation vers une source détectée et pour son suivi durant les mouvements du robot. Yan *et al.* (2011) se focalisent sur l'association d'un stimulus auditif à une source visuelle lorsque plusieurs sources sont visibles, par une approche probabiliste là encore. Une carte audiovisuelle est ainsi apprise pour une source visuelle unique sur la base de critères temporels et spatiaux, et les auteurs démontrent que la carte est en mesure de s'adapter à des configurations non apprises incluant plusieurs sources visuelles. Une expérimentation sur une tête robotique permet de valider cette approche dans un contexte d'interaction homme-robot. Selon la configuration de l'environnement, l'une ou l'autre des modalités n'est pas forcément pertinente, il alors possible de privilégier la modalité la plus adéquate au contexte, comme le montrent Khalidov *et al.* (2008, 2011) ou encore Alameda-Pineda *et al.* (2011) en proposant un modèle probabiliste à base de GMM.

3.3.2 Reconnaissance et séparation de sources

La reconnaissance de source, en particulier la reconnaissance de la parole, peut évidemment s'effectuer à partir d'un signal monaural. Cependant Breteau *et al.* (2010) et Youssef *et al.* (2010) ont démontré que l'utilisation d'un signal binaural apporte plus de robustesse dans cette tâche. Néanmoins lorsque plusieurs sources distractrices sont en présence, les performances de reconnaissance décroissent rapidement comme nous l'avons décrit au paragraphe 2.1.2 en évoquant l'effet « cocktail party », si bien qu'il est nécessaire d'ajouter une phase de séparation permettant d'isoler la source d'intérêt de ses distracteurs. Ainsi Lyon (1983) propose un modèle pionnier en la matière, qui est à la base de nombre d'implémentations robotiques actuelles. L'ITD est d'abord calculé à partir d'une paire de bancs de filtres cochléaires et d'un calcul de corrélation. Les pics d'ITD permettent alors de segmenter le cochléogramme en fonction de la direction estimée à chaque point du plan délai-fréquence. Cette segmentation du cochléogramme en différentes sources est à la base de l'estimation d'un masque binaire permettant d'isoler une source cible des bruits interférants. Roman *et al.* (2003) proposent ainsi une méthode d'estimation de masques binaires, basée sur un apprentissage supervisé, qui permet une amélioration significative du taux de reconnaissance de parole dans une configuration multisources. Des méthodes similaires ont également été proposées par Kim *et al.* (2006) et Park (2006) notamment. En plus des indices spatiaux, Wrigley & Brown (2007) ajoutent à leur modèle de séparation de sources de parole un indice fréquentiel, plus précisément une estimation de la fréquence fondamentale de chaque source. Weiss *et al.* (2011) proposent quant à eux de prendre en compte des informations de haut-niveau relative à l'identification du locuteur, l'algorithme permet une amélioration conséquente des performances, jusqu'à une augmentation de 2.7 dB du SNR, mais demeure néanmoins très sensible aux conditions d'apprentissage.

3.4 Discussion

Les méthodes consacrées à la localisation binaurale sont nombreuses et variées, comme l'a montré ce chapitre. Cette discussion revient tout d'abord sur les méthodes à base d'apprentissage, puis sur les modèles de perception active. Ceci permettra finalement de motiver l'utilisation de la théorie sensorimotrice dans le cadre de la localisation de sources sonores.

Apprentissage Bien que les solutions proposées aujourd'hui semblent tout à fait adaptées à un environnement contrôlé et connu à l'avance, le défi principal de ce domaine de recherche reste l'amélioration des capacités de localisation dans des environnements complexes, dynamiques et, surtout, inconnus et donc non modélisables *a priori*. Pour ce faire, des méthodes d'apprentissage ont été proposées, comme décrit au paragraphe 3.2.4, et utilisées à différentes fins : apprentissage de HRTF, prise en compte de la dynamique de la source ou encore apprentissage d'une mise en correspondance de l'espace des indices binauraux avec une direction dans l'espace physique. Mais quelle que soit la méthode utilisée ou la finalité de cet apprentissage, le « prix » à payer est toujours le même. L'ajout de connaissances *a priori* entraîne en effet une perte de généralité du modèle dès lors que le contexte environnemental s'écarte de celui pris en compte lors de l'apprentissage. Ceci entraîne également bien souvent un accroissement du nombre des paramètres, et donc de la complexité intrinsèque du modèle. Les différents *a priori* pouvant être pris en compte sont là encore très variés : distance interaurale, HRTF, acoustique de l'environnement, nombre de sources actives en sont autant d'exemples. Malgré leur variabilité, ces différents modèles héritent d'une vision *bottom-up* de la perception, dans laquelle chaque bloc de traitement demeure indépendant et ignorant des traitements en amont et en aval et où, si une connexion efférente est présente, elle a bien généralement pour origine les niveaux les plus élevés et les plus symboliques du modèle. Or nous avons vu au chapitre 2 que cette organisation de la perception ne correspond en aucun cas à l'organisation biologique du système auditif. Celui-ci s'illustre en effet par la densité des connexions efférentes à tous les niveaux de traitement, y compris aux niveaux les plus périphériques. Slaney (1997) revient sur cette différence fondamentale entre la réalité biologique et les modèles d'audition artificielle actuellement proposés, en s'inspirant de constats similaires concernant la vision par ordinateur, et identifie la compréhension, la conception et l'intégration de tels systèmes efférents comme un enjeu majeur en sciences de la perception.

Audition active C'est précisément dans ce contexte qu'il faut inscrire les modèles d'audition active décrits au paragraphe 3.2.3. Ces modèles, dont les véhicules de Braitenberg (1986) constituent un bon exemple, proposent ainsi une approche exploitant des boucles sensorimotrices intervenant aux plus bas niveaux de la perception, l'interaction directe entre l'action et la perception permettant alors de s'affranchir de toute représentation ou connaissance de haut niveau requise par les méthodes plus classiques. Cette absence de tout modèle interne contraint cependant le robot à des comportements réactifs où toute perception strictement passive est impossible. En plus de cette différence d'approche entre ces deux classes de modèles, nous notons également une différence dans la finalité recherchée. Ainsi les modèles principalement actifs recherchent bien souvent à accomplir une tâche précise, directement associée à la « survie » du robot, s'approcher ou s'éloigner d'une source sonore par exemple,

mais sans s'intéresser directement à la valeur de l'angle d'incidence de cette source. A l'inverse les modèles associés à l'approche passive recherchent généralement l'estimation la plus précise possible de cet angle, indépendamment d'une tâche ou d'un intérêt écologique particulier.

Approche sensorimotrice En cherchant à estimer l'angle d'incidence d'une source sonore à partir d'indices auditifs, les approches *bottom-up* font donc certaines hypothèses sur les propriétés de l'espace physique – espace tridimensionnel, homogène, isotrope, etc – et sur l'interaction entre le robot et son environnement – structure des HRTF par exemple. Par ce travail de modélisation elles apportent donc au robot une connaissance implicite du monde extérieur. En mettant en relation le retard interaural avec un angle dans l'espace physique, l'Eq. 3.3 suppose ainsi une géométrie euclidienne et une tête acoustiquement transparente. A l'inverse, le point de vue soutenu par la théorie des contingences sensorimotrices est que nos représentations internes sont non pas données *a priori* mais acquises grâce à l'expérience sensorimotrice, c'est-à-dire par l'analyse des interactions entre mouvement effectué et perception ressentie. Basé sur cette théorie, les chapitres 5 et suivants proposent ainsi un modèle de la localisation permettant l'apprentissage autonome de cette tâche par un robot initialement naïf. Nous n'entrerons pas ici dans les détails puisque le chapitre 5 est entièrement consacré à cette approche sensorimotrice de la perception, tant du point de vue théorique qu'applicatif. Notons enfin que nous avons volontairement omis de parler des quelques modèles auditifs existants qui reposent sur la théorie des contingences sensorimotrices. Ceux-ci seront décris en détail au paragraphe 5.1.3.

Chapitre 4

Système auditif artificiel bioinspiré

Sommaire

Nous avons vu au chapitre précédent que les modèles de localisation de sources sonores, qu'ils soient ou non appliqués à la robotique, reposent sur des traitements auditifs permettant d'extraire les indices nécessaires à la localisation. La présentation du modèle auditif développé durant cette thèse est précisément l'objet de ce chapitre. Notre modèle, qui permet d'extraire les indices d'ILD et d'ITD nécessaires à la localisation de sources sonores à partir d'un capteur binaural, se place à mi-chemin entre les modèles neurocomputationnels, visant à modéliser et comprendre les mécanismes biologiques sous-jacents à la localisation, et les modèles « ingénieur-centrés » recherchant avant tout performance et efficacité dans un cadre applicatif bien défini. En effet notre modèle, dont un schéma introductif est proposé Fig. 4.1,

reproduit globalement l'organisation biologique du système auditif jusqu'au niveau sous-thalamique, chaque élément le constituant - oreille externe, modèle de cochlée, etc. - s'inspirant directement de son équivalent biologique tout en garantissant une exécution temps réel du modèle sur une plateforme robotique. L'objectif n'est en effet pas de reproduire le plus fidèlement possible les caractéristiques du système auditif du mammifère mais au contraire de proposer un système simple, efficace et robuste pouvant être utilisé dans le contexte de la robotique autonome et de la perception active.

Ainsi ce chapitre décrit en détail les différents éléments du modèle et se décompose en trois parties. Premièrement nous décrivons les éléments associés au système auditif périphérique : l'oreille externe, la cochlée et la transduction auditive. Le modèle de transduction que nous proposons, basé sur une représentation impulsionnelle de l'information auditive, constitue un aspect original et tout à fait central de notre modèle puisque les traitements en aval se baseront de manière efficace sur cette représentation. La seconde partie de ce chapitre se concentre sur l'extraction des indices binauraux utiles à la localisation. Nous proposerons ainsi une implémentation originale d'un modèle d'extraction de la différence interaurale de temps d'arrivée, qui constitue un classique de la littérature, et nous verrons que cette implémentation s'avère particulièrement adaptée à un contexte robotique. Enfin, la troisième et dernière partie de ce chapitre propose une série d'expérimentations simples effectuées dans un contexte purement passif, qu'elles soient effectuées à partir d'un enregistrement binaural ou en simulation, pour différents contextes acoustiques et différentes sources sonores. Ces expérimentations permettent de valider notre modèle en situation et d'en dégager les propriétés importantes, avant son utilisation de manière active et « incarnée » aux chapitres suivants. Nous verrons également que ce modèle est en mesure de reproduire qualitativement quelques propriétés du système auditif bien connues en psychoacoustique pour leur contribution à la localisation de sources.

4.1 Système auditif périphérique

Le système auditif périphérique présenté ici se compose d'une paire de pavillons artificiels incluant chacun un microphone, d'une paire de cochlée et d'un modèle de transduction auditive permettant une représentation impulsionnelle de l'information, à la manière de l'information afférente circulant sur le nerf auditif. Ses différents composants sont décrits en détail dans les paragraphes suivants et accompagnés à chaque fois d'un état de l'art de la littérature correspondante.

4.1.1 Modèles d'oreille externe

L'oreille externe est le composant le plus périphérique du système auditif, il est en contact physique avec l'environnement. Le filtrage opéré par l'oreille externe est ainsi dépendant du contexte dans lequel le modèle auditif est utilisé. Les résultats issus de cette thèse exploitent 4 différentes modélisations d'oreilles externes que nous décrivons ci-dessous, les 2 premières correspondant à un modèle physique et les 2 dernières à une simulation.

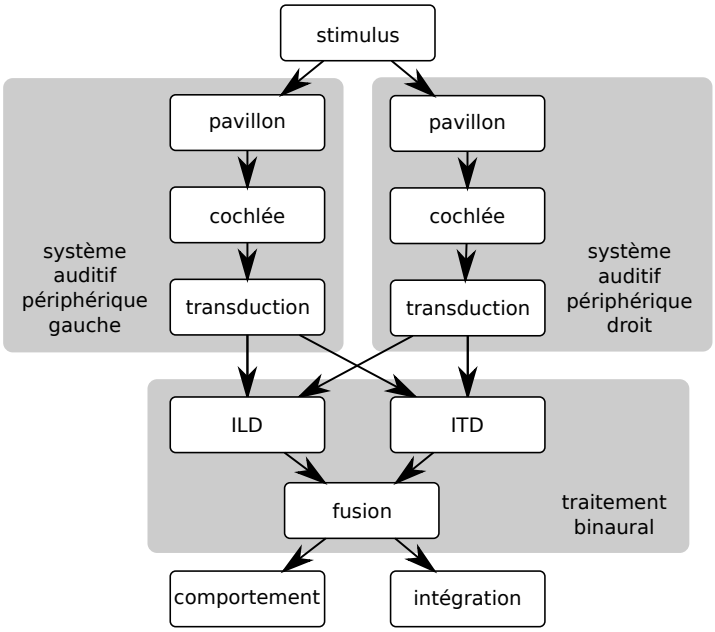


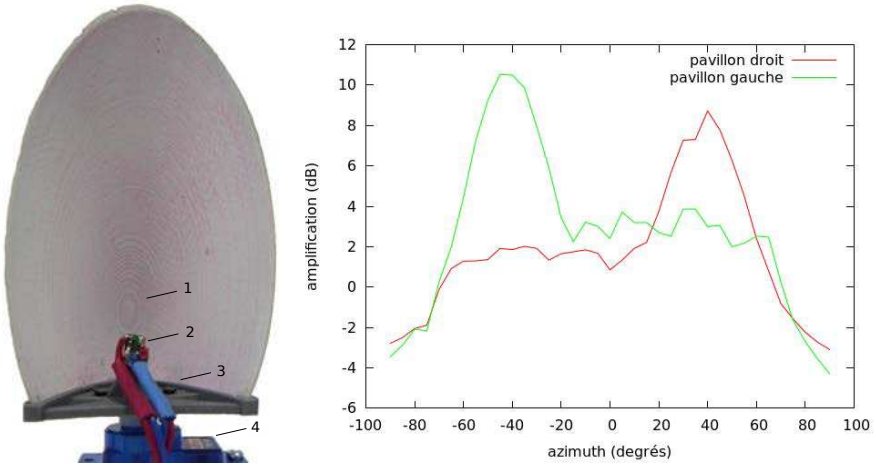
Figure 4.1 – Schéma du modèle de système auditif binaural. Il est composé d'une paire de systèmes périphériques indépendants l'un de l'autre dont la sortie est connectée à un étage de traitement binaural. La localisation proprement dite est effectuée par des comportements réactifs et une phase d'intégration sensorimotrice introduits aux chapitres 5 et suivants.

4.1.1.1 Pavillons du robot-rat Psikharpax

La plate-forme Psikharpax (Meyer *et al.* , 2005; Caluwaerts *et al.* , 2012) est équipée d'une paire de pavillons artificiels[1]. Chaque oreille externe, dont un exemplaire est présenté Fig. 4.2(a), se compose d'un pavillon parabolique en plastique monté sur un support, lui-même fixé sur un servomoteur relié à la tête du robot. Ce servomoteur procure un degré de liberté à chaque pavillon indépendamment et leur permet une rotation dans le plan azimutal. Un microphone MEMS, placé dans une encoche du support, est dirigé vers le centre de la paraboloïde. Une fois réfléchi par le pavillon, le signal capturé par le microphone est numérisé par une carte son Terratec Aueron USB embarquée sur le robot. L'échantillonnage est fixé à 44.1 Hz pour une résolution de 16 bits. Les signaux des pavillons gauche et droit sont fusionnés en un signal stéréo.

Le centre de la paraboloïde, vers lequel est orienté le microphone, jouent le rôle d'une "fovéa auditive" et permet d'amplifier le son en provenance de l'avant du pavillon. On constate ainsi sur la Fig. 4.2(b) que la géométrie du pavillon et l'orientation du microphone permettent une amplification de 10 dB environ pour une source localisée en face du pavillon. Ce filtrage, enrichi par le degré de liberté en rotation, reproduit ainsi les caractéristiques fondamentales de l'oreille externe du mammifère, tel que précisé dans le paragraphe 2.2.1. Il est à noter que les propriétés acoustiques précises des pavillons du robot-rat n'ont pas été mesurées expérimentalement, les HRTF des deux oreilles externes une fois montées sur le robot ne sont donc pas connues. Néanmoins le rôle fondamental de l'amplification causée par le pavillon

1. Ces pavillons ont été conçus et réalisés en prototypage rapide par Christophe Grand, Maître de Conférence à l'Institut des Systèmes Intelligents et de Robotique.



(a) Pavillon artificiel (b) Amplification des pavillons

FIGURE 4.2 – (a) Pavillon artificiel du robot-rat Psikharpax. 1 - fovéa auditive. 2 - microphone. 3 - support. 4 - servomoteur. (b) Amplification des pavillons gauche et droit par rapport à une utilisation sans pavillons, les pavillons gauche et droit sont orientés à $-45°$ et $45°$ respectivement (dans le sens anti-trigonométrique), le signal émis est un bruit blanc positionné en face du robot à 1 m de distance environ. Les différents azimuts sont obtenus par rotation du cou.

dans une tâche de localisation active sera démontré au paragraphe 6.1.2. Les données concernant les HRTF des oreilles externes ne sont de plus pas nécessaires au modèle de localisation basé sur l'approche sensorimotrice. Ce modèle, introduit au paragraphe 5.3, permet en effet l'apprentissage implicite de cette fonction de transfert au travers de l'apprentissage des interactions sensorimotrices.

4.1.1.2 Mannequin binaural

Un mannequin binaural est parfois utilisé pour l'étude et la modélisation du système auditif. Ce type de mannequin, dont on peut voir une illustration Fig. 3.1(c), reproduit en fait les caractéristiques acoustiques d'une tête humaine. Deux microphones sont installés à l'emplacement des tympans et sont entourés de pavillons artificiels reproduisant finement la morphologie du pavillon humain. Nous utiliserons directement des enregistrements effectués sur ce type de mannequin dans le chapitre 5. Il s'agit de la base de données CAMIL (*Computational Audio-Motor Integration through Learning*), proposée par Deleforge & Horaud (2011). Pour une description complète de cette base de données, se référer au paragraphe 5.3.1.1.

4.1.1.3 Simulation de HRTF

L'oreille externe peut être également modélisée non pas directement par un modèle physique mais par l'intermédiaire de ses HRTF. Il est ainsi possible de reconstruire les HRTF à partir de données géométriques (Lopez-Poveda & Meddis, 1996; Algazi *et al.* , 2002) ou anthropométriques (Satarzadeh *et al.* , 2007; Rodemann, 2011). Cette modélisation est facilitée par le fait que le torse, la tête et les pavillons agissent de manière indépendante et additive (Algazi *et al.* , 2001a). Une autre approche est basée sur l'enregistrement de ces HRTF sur des sujets humains ou des

pavillons artificiels. Différentes bases de données sont ainsi disponibles, concernant la localisation en azimut et en élévation (Gardner, 1994; Algazi *et al.* , 2001b) mais également en distance (Wierstorf *et al.* , 2011). Nous utiliserons directement cette méthode et les données proposées par Algazi *et al.* (2001b) lors d'expérimentations effectuées en simulation et décrites dans le paragraphe 5.3. Ainsi, en considérant un signal temporel $x_e(t)$ émis par une source sonore donnée puis filtré par l'oreille externe, le signal $x_s(t)$ obtenu au niveau du tympan peut être exprimé comme :

$$x_s(t) = \frac{h(\theta, \phi) * x_e(t)}{d}, \tag{4.1}$$

où $h(\theta, \phi)$ est la réponse impulsionnelle correspondant à une source positionnée à l'azimut θ, à l'élévation ϕ et à la distance d par rapport au centre de l'axe binaural. $*$ est l'opérateur de convolution.

Les enregistrements de HRTF sont classiquement effectués à un nombre restreint de positions. Algazi *et al.* (2001b) par exemple proposent un pas de 5° mais des méthodes d'interpolation permettent de palier cette limitation et de simuler une source sonore perçue à une position proche des positions enregistrées, à base notamment d'interpolations de splines ou d'interpolation disjointe des indices d'ILD et d'ITD (Larcher & Jot, 1997; Cheng & Wakefield, 2001; Corey & Wakefield, 2001). Nous utiliserons ici une technique d'interpolation par pondération inverse à la distance (Shepard, 1968) à partir de réponses impulsionnelles enregistrées sur un mannequin binaural (Algazi *et al.* , 2001b). Considérons ainsi la base de données $\mathcal{H}$ composée de m fois n réponses impulsionnelles enregistrées à m différents azimuts et n différentes élévations, de sorte que $\mathcal{H} = \{h_{ij}(\theta_i, \phi_j) | i \in [1, m], j \in [1, n]\}$. L'interpolation d'une réponse impulsionnelle $\tilde{h}(\theta, \phi)$ pour un azimut θ et une élévation ϕ quelconque est obtenue à partir de ses k plus proches voisins (k-ppv) dans $\mathcal{H}$, c'est-à-dire les k éléments $h_{(k)}(\phi_{(k)}, \theta_{(k)})$ de $\mathcal{H}$ minimisant la distance angulaire $(\phi - \phi_{(k)})^2 + (\theta - \theta_{(k)})^2$. L'interpolation de $\tilde{h}(\theta, \phi)$ s'exprime finalement comme une somme pondérée des $h_{(k)}$:

$$\tilde{h}(\theta, \phi) = \sum_{i=1}^{k} \frac{w_i h_{(k)}}{\sum_{j=1}^{k} w_j}, \text{ avec } w_i = \frac{1}{\sqrt{(\theta - \theta_{(k)})^2 + (\phi - \phi_{(k)})^2}}. \tag{4.2}$$

En pratique seuls les quatre plus proches voisins sont considérés pour une interpolation 2D, deux voisins seulement pour une interpolation 1D (par exemple une interpolation en azimut uniquement). Cette méthode d'interpolation inverse à la distance (il s'agit dans le cas ci-dessus d'une différence angulaire) est également utilisée au chapitre 5 à des fins d'intégration sensorimotrice (voir l'Eq. 5.4).

4.1.1.4 Filtrage directionnel

Nous présentons enfin un dernier modèle d'oreilles externes pour lequel est uniquement simulée la directivité azimutale, indépendemment de la fréquence. Ce modèle entraîne une simple modulation de l'amplitude du signal d'entrée $x_e(t)$ selon une fonction $h(\phi)$ de l'azimut de la source. Considérant donc un angle $\phi \in [-\frac{\pi}{2}, \frac{3\pi}{2}[$ en radian cette fonction, constituée d'une combinaison de fonctions gaussiennes, s'exprime comme :

$$h(\phi) = \begin{cases} e^{-\frac{(\phi - \phi_\mu)^2}{2\phi_\sigma^2}} & \text{si } -\frac{\pi}{2} \leq \phi < \frac{\pi}{2} \\ 2h(0) - h(\phi - \pi) & \text{si } \frac{\pi}{2} \leq \phi < \frac{3\pi}{2} \end{cases} \tag{4.3}$$

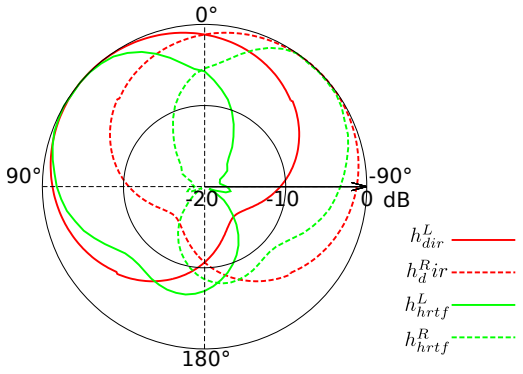


FIGURE 4.3 – Moyenne quadratique normalisée de la réponse des deux filtres d'oreille externe h_{hrtf} (modèle 3) et h_{dir} (modèle 4) (en dB) en fonction de l'azimut ϕ de la source par rapport à l'axe binaural (en deg.) en réponse à un bruit blanc. La direction frontale correspond à $\phi = 0°$.

avec $\phi_\mu = 0$ la moyenne de la gaussienne, correspondant à l'angle de directivité maximale, et $\phi_\sigma = \frac{\pi}{2}$ sa variance. Le choix de cette fonction assure une pondération du signal d'entrée par un coefficient compris dans l'intervalle $[0, 1]$. Les oreilles externes gauche et droite sont modélisée par un décalage de la fonction $h(\phi)$ de $\frac{\pi}{4}$ et $-\frac{\pi}{4}$ respectivement. Les signaux gauche et droit $x^l(t)$ et $x^r(t)$ s'expriment donc pour un azimut ϕ en fonction du signal d'entrée $x^e(t)$ comme :

$$x^l(t) = h(\phi + \frac{\pi}{4})x^e(t) \text{ et } x^r(t) = h(\phi - \frac{\pi}{4})x^e(t) \tag{4.4}$$

Bien qu'ils différent selon la présence ou l'absence d'indices spectraux, ces deux modèles précédents ont en commun une directivité azimutale comparable, comme l'illustre la Fig. 4.3. La comparaison de résultats de localisation obtenu grâce à ces deux modèles nous permettra d'une part de montrer que les indices directionnels sont suffisant pour résoudre l'ambiguïté avant/arrière de manière active (voir le paragraphe 6.1.2), mais que la présence d'indices spectraux est requise pour sa résolution de manière passive (voir le paragraphe 6.1.1).

4.1.2 Filtrage cochléaire

Les CCI disposées dans la cochlée, le long de la membrane basilaire, sont chacune sensible à une fréquence caractéristique différente. La tonotopie ainsi créée est l'aspect primordial que les différents modèles de cochlée cherchent à reproduire, chaque élément du banc de filtre modélisant la réponse de la membrane basilaire autour d'un point précis. Il existe aujourd'hui une grande variété de modèles cochléaires, allant des simples modèles linéaires et passifs à des modèles analytiques et physiques beaucoup plus complexes (Lyon *et al.* , 2010; Fay & Popper, 2010). Ils varient également selon leurs applications, incluant historiquement le traitement de la parole (Warren, 2008), mais aussi les implants cochléaires (Moore, 2003; Wilson *et al.* , 1994; Wilson & Dorman, 2008), la modélisation en mécanique cochléaire (Beyer, 1992; Givelberg & Bunn, 2003; Givelberg *et al.* , 2001) et bien sur la robotique (van Schaik & Shamma, 2004; Chan *et al.* , 2007; Liu *et al.* , 2010).

Les modèles dédiés à la parole ou à la robotique admettent par nécessité une description simple et permettent des applications embarquées et/ou temps-réel (Brucke

et al. , 1998, 1999; Leong *et al.* , 2003; Katsiamis *et al.* , 2009). En première approximation ces modèles à base de bancs de filtres peuvent être divisés en deux groupes : cascade et parallèle (Lyon *et al.* , 2010). Les modèles cascade reproduisent la propagation de l'onde acoustique le long de la membrane basilaire par filtrages successifs, chaque nouveau filtre ajoutant son effet aux filtres précédents. Dans les modèles parallèles les filtres sont au contraire indépendants les uns des autres. La réponse locale de la membrane est reproduite par filtrage passe-bande et, bien que ces filtres nécessitent un ordre plus élevé que les filtres cascades, ils permettent en général une implémentation plus aisée (Slaney, 1993; Fay & Popper, 2010).

Deux modèles différents ont été utilisés durant cette thèse et sont présentés ci-dessous : un modèle cascade appelé modèle de Lyon et un modèle parallèle à base de filtres gammatones. Les filtres gammatones seront par la suite utilisés dans toutes les expérimentations concernant la modalité auditive, à l'exception des expériences concernant la phonotaxie présentées au paragraphe 6.1.2 qui utilisent le modèle de Lyon. Les filtres gammatones sont également à la base du modèle de banc de vibrisses introduit dans l'annexe B.

4.1.2.1 Modèle de Lyon

Le modèle de Lyon (1982), schématisé sur la Fig. 4.4, se propose de reproduire l'activité des fibres du nerf auditif en utilisant un degré de modélisation particulièrement intéressant pour un usage en robotique. La membrane basilaire est ainsi vue comme une cascade de filtres du second ordre. Chacun de ces filtres est composé d'un filtre coupe-bande et d'un résonnateur. Les coupe-bandes, disposés en cascade, opèrent à des fréquences de plus en plus basses et modélisent ainsi la propagation de l'onde acoustique le long de la membrane basilaire, de la base vers l'apex. Les résonnateurs permettent quant à eux de reproduire la sélectivité fréquentielle en un point précis de la membrane. Pour chaque association d'un coupe-bande avec un résonnateur, la fréquence de coupure du premier est alignée sur la fréquence de résonance du second. Les fonctions de transfert de ce filtrage cumulé coupe-bande/passe-bande est illustré Fig. 4.5(a) et un cochléogramme obtenu à partir de ce modèle est illustré Fig. 4.6(b). En plus de cet étage de filtrage le modèle de Lyon se compose d'un étage de rectification de demi-onde (HWR) et d'un étage de contrôle automatique de gain (AGC). La HWR consiste en pratique à ne conserver que les valeurs positives du signal en sortie du filtre, reproduisant de la sorte le caractère unidirectionnel du fonctionnement des CCI. Les canaux ioniques de ces dernières ne s'ouvrent en effet que lors d'une déflexion positive de la membrane basilaire. L'AGC permet enfin de réduire la dynamique du signal d'entrée pour l'adapter à celle, plus limitée, du nerf auditif (Lyon, 1990). Ce modèle de cochlée a été implémenté en C++ par la société Brain Vision Systems sur la base de l'implémentation proposée par Slaney (1988).

4.1.2.2 Modèle gammatone

A l'inverse du modèle de Lyon, les bancs de filtres gammatone ont une architecture purement parallèle. Ils ont été introduits notamment par Patterson *et al.* (1987, 2003) pour la modélisation de la membrane basilaire et sont aujourd'hui largement utilisés comme filtres cochléaires (Patterson *et al.* , 1992; Slaney, 1993), la réponse d'un filtre gammatone reproduit en effet fidèlement la réponse linéaire de la membrane basilaire en un point donné (Patterson *et al.* , 1987; Glasberg & Moore, 1990). De nombreux modèles non-linéaires à base de filtres gammatones ont également été

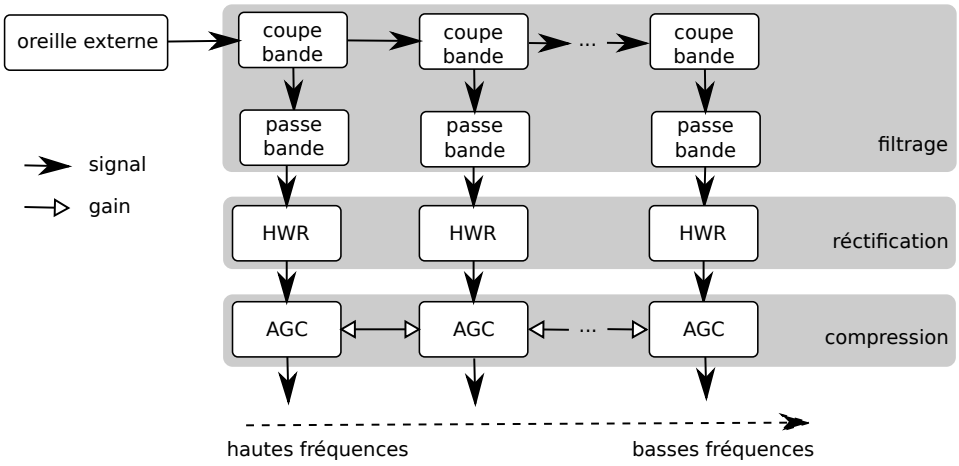


FIGURE 4.4 – Modèle de Lyon (1982).

proposés pour affiner le degré de modélisation (Katsiamis *et al.* , 2007), tels que les filtres *all pole* gammatone (Lyon, 1996) et les filtres gammachirp (Irino & Patterson, 1997, 2001, 2006). Ces modèles sont généralement composés d'une voie linéaire et d'une voie non-linéaire dont les réponses s'ajoutent en sortie.

L'approche gammatone décompose donc la membrane basilaire en plusieurs filtres indépendants les uns des autres. La bande passante de chacun de ces filtres est représentée par une bande passante rectangulaire équivalente (ERB) dont la largeur augmente avec la fréquence et, considérant un filtre à la fréquence caractéristique f_c, Glasberg & Moore (1990) proposent l'équation suivante pour le calcul de l'ERB :

$$ERB(f_c) = 24.7(\frac{4.37 f_c}{1000} + 1). \tag{4.5}$$

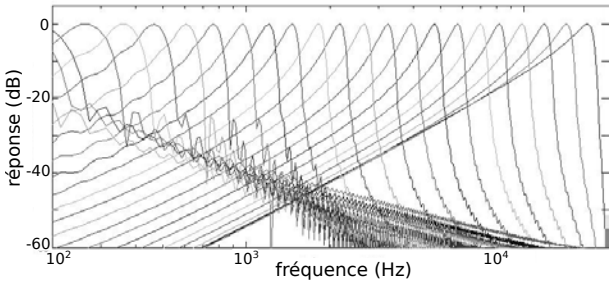
Suivant les notations de Slaney (1993), cette équation peut également être réécrite et généralisée comme :

$$ERB(f_c) = (\frac{f_c}{ear_Q}^n + {min_{BW}}^n)^{1/n}, \tag{4.6}$$

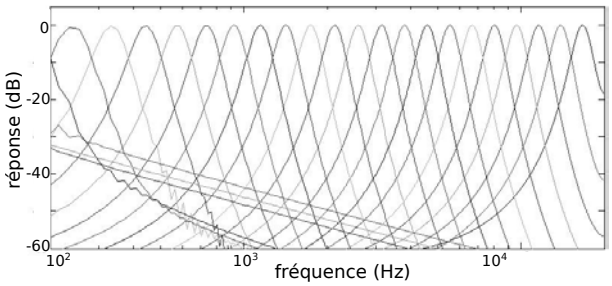
où ear_Q est la réponse asymptotique du filtre en hautes fréquences, min_{BW} est la bande passante minimale dans les basses fréquences et n est l'ordre du filtre. Nous utilisons ici les paramètres suggérés par Glasberg & Moore (1990) qui correspondent à l'équation 4.5, c'est-à-dire $ear_Q = 9.26$, $min_{BW} = 24.7$ et $n = 1$. Il est néanmoins utile d'introduire ces précisions sur le calcul de la bande passante car nous serons amenés à modifier ces valeurs "par défaut" réglées sur l'audition humaine pour modéliser un banc de vibrisses dans l'annexe B. Une fois la bande passante d'un filtre définie en fonction de sa fréquence caractéristique, il s'agit maintenant de calculer la fréquence caractéristique de chaque filtre du banc. Ainsi, étant donné un banc de filtre composé de n canaux et aux fréquences caractéristiques minimale f_l et maximale f_h, Slaney (1993) propose de calculer la fréquence caractéristique f_c^i du filtre $i \in [1, n]$ comme ceci :

$$f_c^i = -\alpha + (f_h + \alpha)e^{-\frac{i}{n}log(f_h+\alpha)+log(f_l+\alpha)}, \text{ avec } \alpha = ear_Q.min_{BW}. \tag{4.7}$$

Avec cette méthode le banc de filtre est donc simplement paramétré par le nombre de filtres le composant, les fréquences caractéristiques minimale et maximale et, dans



(a) Modèle de Lyon



(b) Filtres gammatones

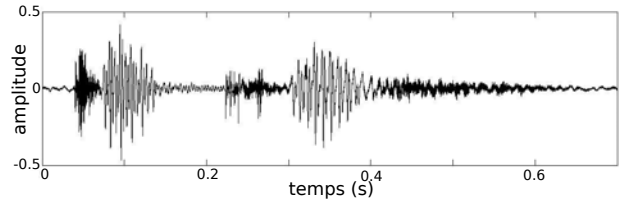
Figure 4.5 – Fonction de transfert des bancs de filtres cochléaires. (a) modèle de Lyon. (b) filtres gammatones. Les 2 bancs de filtres comptent 25 canaux entre 20 Hz et 20 kHz et les gains sont normalisés. D'après Slaney (1998).

le cas d'une implémentation numérique, la fréquence d'échantillonage du signal en entrée. Un filtre gammatone est construit comme la combinaison d'une loi gamma et d'une sinusoïde, sa réponse impulsionnelle $g(t)$ s'exprimant dans le domaine temporel comme :

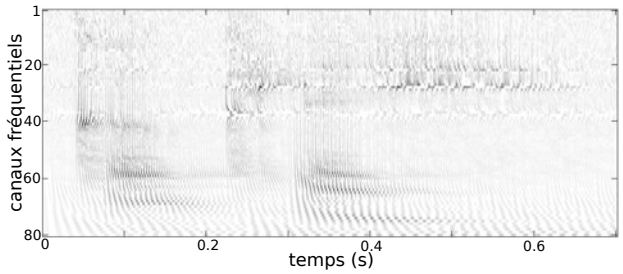
$$g(t) = at^{n-1}e^{-2\pi bt}cos(2\pi f_c t + \Phi), \tag{4.8}$$

où a défini le gain du filtre, b sa bande passante, f_c sa fréquence caractéristique et Φ sa phase, avec $t > 0$. Patterson *et al.* (1992) proposent $b = 1.019ERB(f_c)$. La fonction de transfert d'un banc de 25 filtres gammatones est présentée Fig. 4.5(b) tandis qu'un cochléogramme obtenu à partir de ce modèle est visible Fig. 4.6(c).

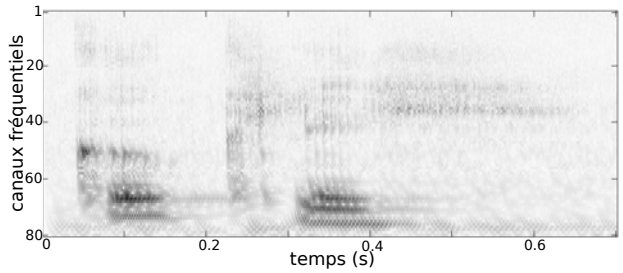
Dans le cadre d'une utilisation robotique, ce filtrage cochléaire est utilisé de manière intensive en temps réel. L'implémentation utilisée ici, elle-même se basant sur le travail de Cooke (1993), est celle proposée par Ma (2006); Ma *et al.* (2007). Nous avons effectué quelques optimisations sur le code original concernant notamment la factorisation des opérations, la gestion de la mémoire et l'intégration en C++. Ces modifications, bien que mineures, sont importantes car elles permettent une amélioration de l'utilisation des ressources. Notons également que, avec les implémentations utilisées, l'exécution du modèle gammatone est environ 3.5 fois plus rapide que pour le modèle de Lyon à conditions égales. Nous avons également porté cet algorithme sur un circuit électronique développé par la société BVS, avec l'objectif de proposer un système de pré-traitement auditif hardware utilisable en robotique. Cette implémentation est détaillée dans l'annexe D.3.



(a) Signal temporel



(b) Modèle de Lyon



(c) Filtres gammatones

FIGURE 4.6 – Cochléogramme du mot "cochlée". (a) signal temporel. (b) modèle de Lyon sans AGC. (c) filtres gammatones. Les 2 bancs de filtres comptent 80 canaux entre 20 Hz et 20 kHz et incluent une étape de HWR. Les données des deux cohléogrammes sont normalisées et compressées à la puissance 1/2.

4.1.3 Représentation impulsionnelle

Une fois que le mouvement de la membrane basilaire a été simulé par un banc de filtre cochléaire, l'étape suivante consiste à modéliser la transduction proprement dite, c'est-à-dire l'action des CCI et la conversion de l'énergie mécanique en train d'impulsions transitant sur le nerf auditif. Bien que le modèle cochléaire de Lyon présenté ci-dessus intègre une phase de HWR et une phase de compression, qui peuvent être vus comme une modélisation en première approximation de la transduction auditive, le signal à la sortie du modèle est un signal continu - bien que discrétisé - et non un train d'impulsions. Parmi les modèles de transduction auditive plus complets, le modèle proposé par Meddis *et al.* (Meddis, 1986; Meddis *et al.* , 1990; Sumner *et al.* , 2002, 2003) fait autorité. La simulation du fonctionnement probabiliste des CCI et de la réponse électrique du nerf auditif permettent en effet aux auteurs de reproduire les caractéristiques temporelle et fréquentielle de réponses évoquées enregistrées sur le nerf auditif du mammifère avec précision.

La représentation impulsionnelle utilisée ici est plus simplement basée sur l'extraction des "pics" des sorties des canaux cochléaires, c'est-à-dire de leurs maxima

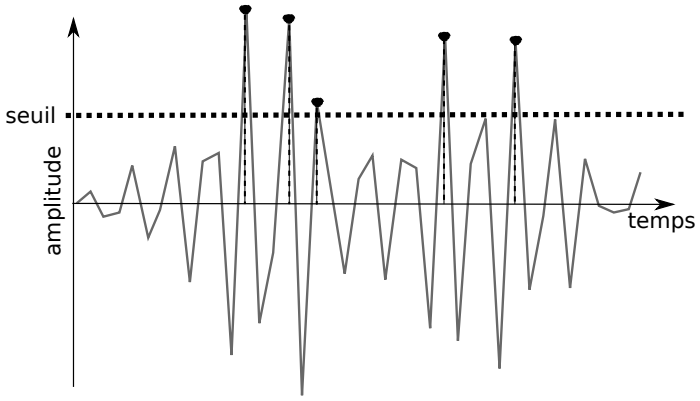


Figure 4.7 – Extraction d'un train d'impulsions depuis la sortie d'un canal cochléaire. Seuls les maxima locaux supérieurs à la valeur du seuil sont conservés.

locaux. Considérant ainsi la sortie $x(t)$ d'un filtre cochléaire, une impulsion $p(t)$ est générée si et seulement si un maximum local est rencontré et si ce maximum est supérieur au seuil d'activité minimale τ. C'est-à-dire pour $\tau \geq 0$:

$$p(t) = \begin{cases} x(t) & \text{si } \frac{dx(t)}{dt} = 0 \text{ et } x(t) > \tau \\ 0 & \text{sinon} \end{cases} \tag{4.9}$$

Par simplicité l'équation 4.9 représente le traitement opéré sur un unique canal cochléaire, comme illustré sur la Fig. 4.7. Bien entendu ce traitement est effectué en parallèle sur chaque canal des filtres gauche et droit.

Cette modélisation de la transduction auditive, proposée par (N'Guyen, 2010; N'Guyen *et al.* , 2011b) pour la transduction du système vibrissal du robot-rat Psikharpax, est motivée par plusieurs arguments. Cela permet tout d'abord une réduction drastique de la quantité d'information, ce qui rend d'autant plus économes les traitements ultérieurs, tout en conservant les caractéristiques du signal en amplitude et en fréquence. Lors de travaux pionniers en psychoacoustique, Licklider & Pollack (1948) ont en effet testé les effets de diverses distorsions du signal sur la capacité de l'humain à reconnaître la parole. Certaines des distorsions testées sont basées sur l'écrêtage infini, une transformation ne conservant que la périodicité du signal dans la mesure où celui-ci est transformé en un signal carré d'amplitude fixe dont le changement de front correspondait au passage du signal source par zéro. En dépit d'une telle réduction d'information, les sujets testés conservaient de très bonnes performances en reconnaissance de la parole. Plus récemment Ghitza (1994) a démontré qu'une telle représentation à base d'impulsions constitue également une base solide pour le traitement artificiel de la parole. D'un point de vue biologique enfin, ce type de code neural semble également être un excellent compromis entre la qualité du codage de signaux auditifs naturels et la complexité - et donc la dépense énergétique - nécessaire à cette opération (Schwartz & Simoncelli, 2001; Lewicki, 2002; Smith & Lewicki, 2006).

Le mécanisme de seuillage a été rajouté au modèle original pour prendre en compte la spécificité de la modalité auditive par rapport à la modalité tactile. Cette dernière est en effet une modalité "de contact" qui est faiblement perturbée par des signaux non désirés, à l'inverse de la modalité auditive qui doit constamment faire face à de multiples signaux distracteurs. Ces signaux peuvent être bruits de fond ou réverbérations et sont souvent de faible intensité relativement à la source sonore

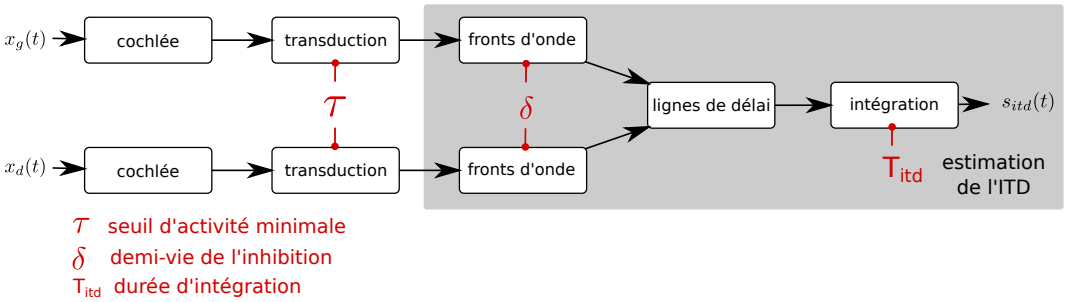


FIGURE 4.8 – Modèle d'extraction de l'ITD. Les fronts d'ondes sont d'abord extraits des deux trains d'impulsions multicanaux en entrée, puis traversent les lignes de délais du modèle de Jeffress. Les coïncidences binaurales détectés sont alors intégrées sur une période T_{itd} et sous-échantillonnées à la fréquence $2/T_{itd}$. Le vecteur d'ITD $s_{itd}(t)$ est ainsi obtenu après intégration. Les paramètres efférents (pouvant être modifiés au cours du traitement) τ, δ et T_{ild}, respectivement associés à la transduction, aux fronts d'ondes et à l'intégration temporelle, sont affichés en rouge.

principale. L'utilisation de ce seuil nous permettra d'inhiber le bruit des moteurs du cou d'un robot lors de nos expérimentations autour du comportement de phonotaxie présenté au paragraphe 6.1.2, nous garantissant que ce comportement n'est pas biaisé par un bruit non désiré. La valeur du seuil est dans ce cas simplement fixée « à la main » pour être supérieure à l'activité de la cochlée durant une rotation du cou dans un environnement silencieux. De plus le seuillage compense également le choix de déclencher une impulsion pour chaque extremum rencontré, ce qui permet de ne pas donner un poids excessif aux intervalles dominés par des fréquences élevées, dont le bruit. Enfin, cette méthode basée sur les maxima locaux permet d'extraire à la fois l'information sur la modulation en amplitude et en fréquence. Chacune de ces deux propriétés du codage impulsionnel sera directement exploité par la suite, l'information sur l'amplitude sera à la base du calcul de l'ILD et l'information temporelle permettra l'extraction de l'ITD.

4.2 Traitements binauraux

Ce paragraphe présente les modèles d'extraction de l'ITD et de l'ILD. Comme nous le verrons, ces deux modèles tirent directement bénéfice de la représentation impulsionnelle décrite ci-dessus. Le paragraphe 3.1 ayant largement fait part de l'importance d'un code efficace dans un contexte robotique, l'implémentation des modèles est également détaillée.

4.2.1 Différence interaurale de temps d'arrivée

Ce paragraphe propose une description complète des traitements liés à l'extraction de l'ITD. Nous présentons tout d'abord un système d'extraction des fronts d'ondes qui, placé en entrée du modèle de Jeffress, permettra d'accroître ses performances de localisation. Nous introduisons ensuite le modèle de Jeffress et l'implémentation que nous proposons de ce modèle pour l'extraction de l'ITD. Une représentation schématique de ce modèle est proposée Fig. 4.8

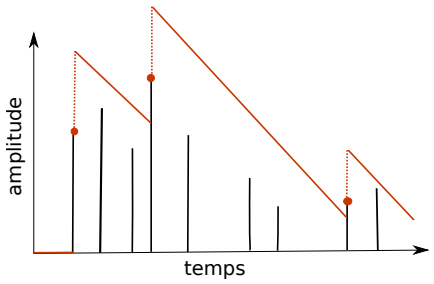


Figure 4.9 – Modèle d'extraction des fronts d'ondes à partir d'un train d'impulsions. Chaque front d'onde détecté (points rouges) correspond à une impulsion dont la valeur est supérieure au seuil dynamique (lignes rouges). Ce seuil est mis à jour en fonction de l'amplitude locale pour chaque front d'onde détecté (pointillés rouges).

4.2.1.1 Extraction des fronts d'onde

L'extraction des fronts d'ondes est un pré-traitement largement utilisé par les divers modèles binauraux existants. L'objectif est d'augmenter la robustesse de la localisation, en présence de bruit ou de réverbération notamment (Smith & Collins, 2007). L'idée générale est qu'un front d'onde correspond au trajet le plus court effectué par un signal acoustique avant d'atteindre l'oreille, correspondant ainsi au signal direct (non-réfléchi) et que, en conséquence, l'information apportée par ce front d'onde est suffisante pour parvenir à une localisation de la source tandis que les signaux secondaires (réfléchis) viennent perturber le processus de localisation. Les fronts d'ondes sont généralement extraits juste avant l'estimation de l'ITD, qui est l'indice binaural le plus sensible aux conditions adverses, mais certains auteurs l'utilisent également pour l'ILD (Heckmann *et al.* , 2006) ou encore, dans un modèle de IC, au moment de la fusion des indices d'ILD et d'ITD (Liu *et al.* , 2009).

Considérant le train d'impulsions $x(t)$ provenant de la transduction d'un unique canal cochléaire, l'extraction d'un front d'onde sur $x(t)$ s'effectue par une simple comparaison avec un seuil dynamique $\tau(t)$ comme l'illustre la Fig. 4.9 . Le train de fronts d'onde $y(t)$ s'exprime alors comme :

$$y(t) = \begin{cases} x(t) & \text{si } x(t) > \tau(t) \\ 0 & \text{sinon} \end{cases} \tag{4.10}$$

Le seuil $\tau(t)$ décroît linéairement entre deux fronts d'ondes et est mis à jour à chaque détection d'un nouveau front. Il est paramétré par son temps de demi-vie δ et par un facteur multiplicatif α, de sorte que :

$$\tau(t) = \begin{cases} \max(0, (1 - \frac{dt}{2\delta})\tau(t - dt)) & \text{si } y(t) = 0 \\ \alpha x(t) & \text{si } y(t) \neq 0 \end{cases} \tag{4.11}$$

Ici $dt = 1/f_s$ correspond à la période d'échantillonage et $\tau(t - dt)$ réfère donc à la valeur du seuil pour l'échantillon précédent. Le temps de demi-vie δ correspond au temps mis par le seuil pour voir sa valeur divisée par 2. Afin d'éviter que ce seuil tombe dans le domaine négatif, sa valeur minimale est fixée à 0. De même nous fixons $\tau(0) = 0$. Le seuil dynamique obéit dans notre modèle à une très simple loi de décroissance linéaire, nous n'avons en effet pas remarqué d'amélioration significative en utilisant une loi exponentielle. Le paragraphe 4.3.2 reviendra sur l'influence des fronts d'onde et du type de seuillage sur l'extraction de l'ITD.

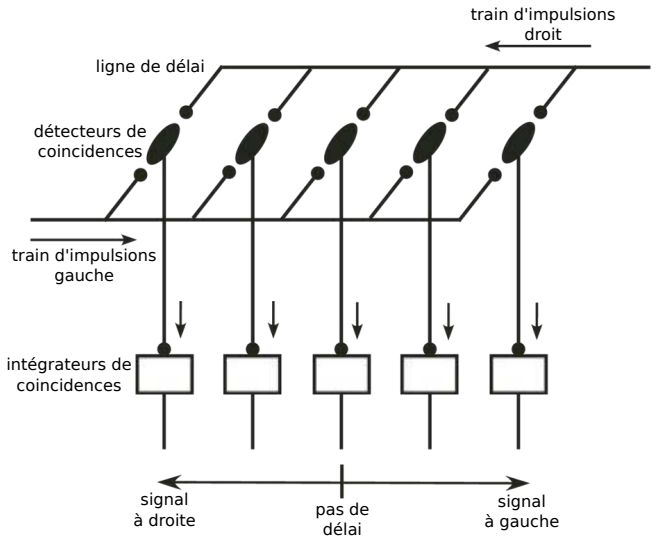


FIGURE 4.10 – Modèle de Jeffress (1948) pour l'extraction de l'ITD à base de lignes de délais. D'après Cariani (2011).

4.2.1.2 Modèle de Jeffress

Le modèle de Jeffress (1948) propose une explication sur la manière dont le système auditif estime la corrélation croisée entre les signaux gauche et droit. Plus précisément, Jeffress propose un mécanisme structurel permettant de représenter spatialement une différence temporelle. Ce modèle, dont une représentation schématique est illustrée Fig. 4.10, reçoit en entrée deux trains d'impulsions, associés aux côtés gauche et droit, contenant des informations temporelles précises sur le signal acoustique sous-jacent. Chaque train d'impulsions traverse une ligne de délai qui permet de retarder artificiellement les impulsions. Des détecteurs de coïncidence permettent ensuite de repérer les impulsions ayant le même retard relatif, ces coïncidences étant ensuite intégrées dans le temps, permettant d'estimer le délai moyen associé à chaque canal fréquentiel (Cariani, 2011). Ce modèle est encore aujourd'hui une source d'inspiration remarquable pour nombre d'études psychoacoustiques et physiologiques et il est virtuellement à la base de tous les modèles d'extraction de l'ITD (Joris *et al.* , 1998; Yin, 2002). Il a notamment été implémenté en version électronique (Lazzaro & Mead, 1989) et utilisé pour la localisation dans un contexte robotique (Calmes *et al.* , 2007).

Plus en détail, ce modèle repose sur trois hypothèses fondamentales : (1) les projections afférentes aux cellules binaurales convoient des informations temporelles précises sur le stimulus auditif ; (2) les cellules binaurales se comportent comme des détecteurs de coïncidence, c'est-à-dire que leur réponse maximale à lieu lorsque les impulsions des côtés ipsilatéral et contralatéral arrivent au même instant ; et enfin (3) les projections ipsilatérales afférentes permettent de recréer artificiellement un délai temporel permettant de compenser le délai naturellement présent dans les projections contralatérales. Ces trois hypothèses ont été confirmées dans le cas des oiseaux, et plus particulièrement chez la chouette effraie, dont les performances en termes de localisation en font un remarquable sujet expérimental. Néanmoins ce modèle ne semble pas expliquer l'encodage de l'ITD chez le mammifère, suscitant un vif débat dans la communauté scientifique (Shamma, 1989; Fitzpatrick, 2002; McAlpine, 2003,

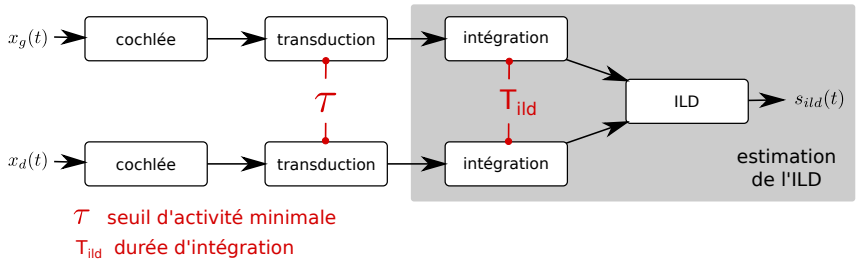


Figure 4.11 – Modèle d'extraction de l'ILD. Les signaux post-transduction, deux trains d'impulsions multicanaux, sont d'abord intégrés sur une période T_{ild} et sous-échantillonnés à la fréquence $2/T_{ild}$. Le vecteur d'ILD $s_{ild}(t)$ est obtenu après intégration comme le ratio normalisé des signaux gauche et droit. Les paramètres efférents τ et T_{ild}, respectivement associés à la transduction et à l'intégration temporelle, sont affichés en rouge.

2005; Joris & Yin, 2007). Le modèle de Jeffress reste néanmoins tout à fait pertinent pour la modélisation bioinspirée du système auditif.

Nous avons vu au paragraphe 3.2.2 que l'ITD permet d'estimer le retard interaural τ_{itd}. Néanmoins, dans le contexte numérique dans lequel nous nous trouvons, ce retard interaural s'exprime non pas directement en secondes mais, en fonction de la fréquence d'échantillonage f_s du signal, en nombre d'échantillons. Le temps de retard est ainsi quantifié par N indices, chacun correspondant à un échantillon de retard par rapport à un délai nul. Nous avons ainsi :

$$N = \lceil f_s \tau_{max} \rceil. \tag{4.12}$$

Finalement la somme des indices à gauche et à droite, en plus de l'indice correspondant à un délai nul, nous donne un total de $2N + 1$ indices différents. Pour un échantillonnage fixé à $f_s = 20$ kHz, cette équation nous donne un total de $2N+1 = 23$ ou $2N + 1 = 29$ délais différents dans l'intervalle $[-90°, 90°]$, considérant les Eq. 3.3 (tête acoustiquement transparente) ou 3.4 (tête sphérique) respectivement.

Soient $n_d = 2N + 1$ le nombre de délais, n_c le nombre de canaux en sortie de cochlée, $s_g(t)$ et $s_d(t)$ les fronts d'onde gauche et droit obtenus par l'Eq. 4.10 à l'instant t, tous deux de dimension n_c. La sortie $s_{itd}(t)$ du modèle de Jeffress s'exprime en fonction de $s_g(t)$ et $s_d(t)$ comme une matrice de dimension n_c par n_d pour laquelle chaque élément (i, j) correspond au nombre de coïncidences détectées pour le canal i et le délai j durant un temps d'intégration T_{itd}. L'implémentation que nous proposons de ce modèle est décrite en détail dans l'annexe D.2.

4.2.2 Différence interaurale d'intensité

Le modèle d'extraction de l'ILD présenté ici reprend le principe de l'Eq. 3.5 exprimant l'ILD comme le ratio des énergies gauche et droite mais, puisque l'entrée est ici un train d'impulsions multicanal, il diffère dans sa mise en oeuvre. Ce modèle est ainsi composé de deux étapes, comme illustré Fig. 4.11. Premièrement, une phase d'intégration temporelle à gauche et à droite permet d'estimer l'intensité moyenne instantanée tout en « lissant » le signal dans le domaine temporel. Une comparaison binaurale des intensités à lieu dans un second temps, l'ILD estimée étant normalisée par rapport à l'intensité moyenne et, en conséquence, indépendante de l'intensité du signal d'entrée. Hartmann & Constan (2002) proposent un modèle d'ILD, basé sur

une intégration temporelle et une estimation de l'ILD indépendante de l'intensité du signal, comme le modèle présenté ici. Ce modèle prédit l'indépendance de l'indice d'ILD par rapport à la corrélation interaurale où à l'intensité, hypothèse vérifiée approximativement par les auteurs dans leur étude psychoacoustique. Voici le détail de ces deux étapes de l'extraction de l'ILD, le calcul étant effectué en parallèle et de manière indépendante sur chacun des canaux fréquentiels.

La phase d'intégration temporelle, monorale, permet donc l'estimation de l'intensité moyenne instantanée à gauche et à droite, respectivement notées $s_g(t)$ et $s_d(t)$, et se calcule donc à partir des deux trains d'impulsions gauche et droit $p_g(t)$ et $p_d(t)$ obtenus par l'Eq. 4.9. Ainsi $s_g(t)$ et $s_d(t)$ s'expriment respectivement comme :

$$s_g(t) = \sum_{u=t-T_{ild}}^{t} p_g(u)^2 \text{ et } s_d(t) = \sum_{u=t-T_{ild}}^{t} p_d(u)^2, \tag{4.13}$$

où T_{ild} correspond à la durée d'intégration. Cette équation présente l'expression mathématique d'une intégration temporelle qui requiert en pratique quelques optimisations. Considérons ainsi le calcul de la somme des $p(t)^2$ de $t - T_{ild}$ à T_{ild}. Dans le cas d'une implémentation naïve se basant sur un tableau, chaque évaluation de l'Eq. 4.13 nécessite un parcours complet du tableau, pour une complexité en $O(T_{ild})$. Une seconde solution, retenue ici, est de remplacer le tableau par une file de type « premier arrivé, premier sorti » et de longueur T_{ild}. Cette file est associée à un accumulateur de sorte que chaque élément ajouté à la file est ajouté à l'accumulateur. De même chaque élément supprimé de la file lui est soustrait. Au prix d'opérations mémoires plus importantes cette simple astuce permet d'accéder à la somme des éléments du tableau en temps constant, le temps de calcul devenant ainsi indépendant de la durée d'intégration. En conséquence de cette implémentation, l'intégration temporelle induit un retard de perception égal à la durée d'intégration T_{ild}. C'est-à-dire qu'un son capturé au temps t ne sera perçu une fois intégré qu'au temps $t + T_{ild}$. De plus les signaux $s_g(t)$ et $s_d(t)$ obtenus par ce calcul sont toujours échantillonnés à la fréquence du signal d'entrée et, en conséquence de l'intégration temporelle, peuvent être sous-échantillonnés sans perte d'information. Afin d'alléger les calculs dans les phases ultérieures du modèle, lors de l'intégration sensorimotrice notamment, les signaux $s_g(t)$ et $s_d(t)$ sont ainsi sous-échantillonnés à la fréquence $f_s = 2/T_{ild}$.

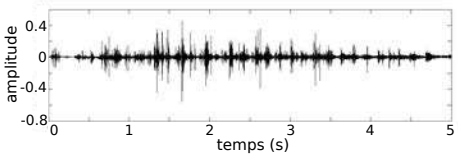
A partir des signaux $s_g(t)$ et $s_d(t)$ intégrés et sous-échantillonnés, l'ILD se calcule finalement comme le rapport suivant à chaque instant t :

$$s_{ild}(t) = \begin{cases} \frac{2s_g(t)}{s_g(t)+s_d(t)} - 1 & \text{si } s_g(t) + s_d(t) \neq 0 \\ 0 & \text{sinon.} \end{cases} \tag{4.14}$$

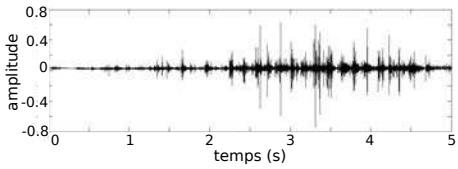
Cette dernière équation exprime donc l'ILD comme un rapport normalisé des intensités gauche et droite, la rendant ainsi indépendante de l'intensité du signal d'entrée. On a donc $s_{ild}(t) \in [-1, 1]$, les trois valeurs caractéristiques que sont -1, 0 et 1 indiquant respectivement une intensité concentrée à droite, centrée ou concentrée à gauche. L'ILD est considérée comme nulle dans le cas d'une intensité sonore nulle, ou *a minima* inférieure au seuil τ associé au calcul de la réponse impulsionnelle.

4.3 Expériences

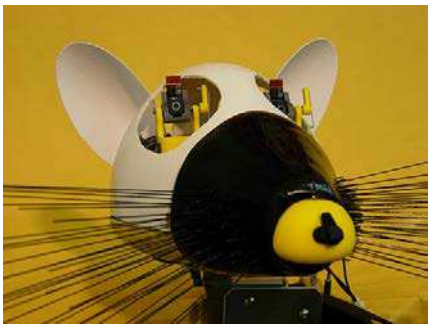
Ce paragraphe décrit plusieurs expériences visant à évaluer le modèle binaural présenté dans ce chapitre dans un contexte purement passif et de confirmer sa capacité à la localisation de sources sonores. Nous étudions tout d'abord les indices d'ILD



(a) Signal temporel gauche.



(b) Signal temporel droit.



(c) Tête de Psikharpax.

Figure 4.12 – Enregistrement binaural obtenu sur la plateforme Psikharpax en conditions passives. Le signal est jeu de clé agité de la gauche vers la droite du robot, de $-90°$ à $90°$ et à 1 m de distance environ. (a)(b) Signaux temporels gauche et droits. (c) Tête de Psikharpax utilisée pour l'enregistrement, équipée de ses vibrisses, d'une paire de caméras, de ses pavillons et de ses microphones.

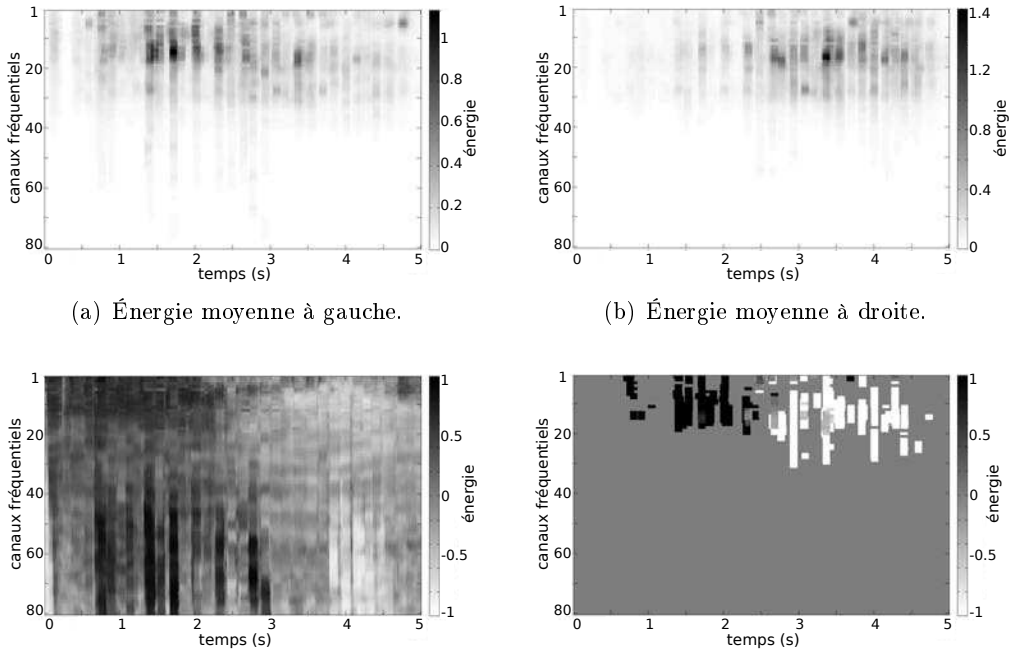
et d'ITD obtenus à partir d'un enregistrement binaural obtenu sur la plateforme Psikharpax en conditions acoustiques réalistes, ce qui nous permettra de valider de manière empirique les aptitudes de notre modèle à la localisation. Nous présentons dans un deuxième temps un jeu de simulations permettant de quantifier plus précisément les performances de localisation. Nous verrons ainsi quel est en pratique le rôle des différents paramètres du modèle. Nous montrons de plus dans l'annexe A que notre implémentation du modèle de Jeffress permet, du moins qualitativement, de reproduire l'effet de précédence.

4.3.1 A partir d'un enregistrement binaural

La Fig. 4.12 présente l'enregistrement d'un signal binaural obtenu à partir de la tête de Psikharpax équipée de ses pavillons. Le signal est jeu de clés agité de la gauche vers la droite du robot, de $-90°$ à $90°$ et à 1 m de distance environ. Ce signal fut enregistré dans un environnement de bureau, incluant bruits de fond (ventilateurs, climatisation) et réverbérations. Le signal, d'une durée 5 s, est largement non-stationnaire. Nous observons de plus une importante amplification à gauche entre 1 s et 2 s, puis à droite entre les secondes 3 et 4. Ceci est dû à l'amplification causée par la directivité des pavillons durant le déplacement de la source (voir Fig. 4.2(b)). Détaillons maintenant les indices binauraux d'ILD puis d'ITD estimés à partir de cet enregistrement.

4.3.1.1 Différence interaurale d'intensité

La Fig. 4.13 représente l'estimation de l'énergie et de l'ILD obtenus. La distribution de l'énergie (Fig. 4.13(a) et 4.13(b)) nous montre que le signal est concentré dans les hautes fréquences, ce qui conformément à la théorie duplex est idéal pour l'estimation de l'ILD. On constate également l'importance du seuil τ, appliqué pendant la transduction, pour le calcul de l'ILD. Puisque l'estimation de cet indice est indépendante de l'intensité, une infime présence d'énergie peut en effet engendrer des valeurs d'ILD extrêmes (Fig. 4.13(c)). L'utilisation du seuil permet ainsi d'inhiber



(a) Énergie moyenne à gauche.

(b) Énergie moyenne à droite.

(c) ILD avec $\tau = 0$.

(d) ILD avec $\tau = 0.1$.

FIGURE 4.13 – Indice d'ILD estimé à partir de l'enregistrement binaural de la Fig. 4.12, avec une durée d'intégration fixée à $T_{ild} = 0.1$ s. (a)(b) Énergies moyennes instantanées à gauche et à droite. (c) ILD estimé avec un seuil $\tau = 0$. (d) ILD estimé avec un seuil $\tau = 0.1$. Le calcul est effectué sur une paire de 80 filtres gammatones disposés dans une bande entre 100 Hz et 8 kHz.

ces composantes de basse énergie et d'assurer que l'estimation de l'ILD n'est effectuée que sur la bande la plus énergétique du signal (Fig. 4.13(d)). L'indice d'ILD s_{ild} estimé par l'Eq. 4.14 ne permet pas d'estimer la direction de la source perçue sans connaissances *a priori* de l'intensité de la source ou des HRTF. Une latéralisation, c'est-à-dire une décision binaire droite ou gauche, est toutefois possible en considérant l'intégration en fréquence de s_{ild}, vecteur qui est composé d'autant de canaux fréquentiels que le banc de filtre cochléaire. Selon son signe, cette somme fournit en effet une information qualitative sur la provenance de la source, comme illustré Fig. 4.14 où un signal d'ILD intégré en fréquence est calculé pour 3 durées d'intégration T_{ild} différentes. On constate qu'une durée d'intégration importante, en plus d'induire un retard dans la perception, inhérent à l'utilisation d'une fenêtre temporelle, permet de stabiliser l'estimation de l'ILD et d'en stabiliser la latéralisation dans le temps.

4.3.1.2 Différence interaurale de temps d'arrivée

À partir de ce même signal, et toujours sur la base d'une paire de 80 filtres gammatone disposés entre 100 Hz et 8 kHz, la Fig. 4.15 illustre les indices d'ITD obtenus pour différents paramètres, nous permettant ainsi de constater l'influence du seuil de transduction τ et du temps de demi-vie δ associé à l'extraction des fronts d'ondes. Constatons tout d'abord que la Fig. 4.15(a), obtenue dans des conditions « neutres » (*i.e.* avec $\tau = 0$ et $\delta = 0$), laisse clairement apparaître un *pattern* en forme de sigmoïde allant d'un délai positif vers un délai négatif, reflétant le déplacement

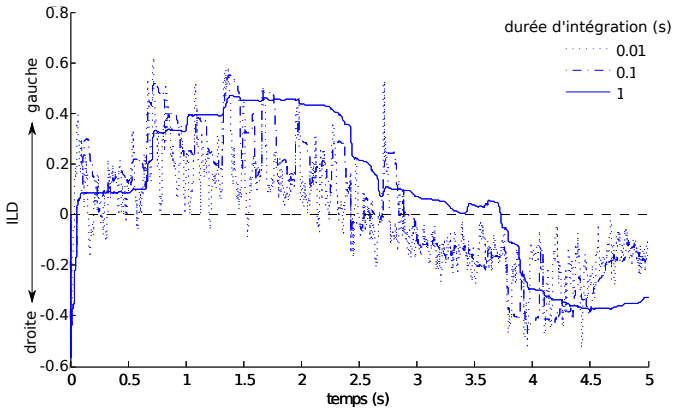
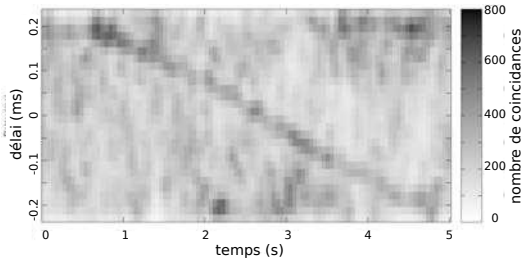


Figure 4.14 – Latéralisation d'une source sonore à partir de l'ILD intégré en fréquence pour 3 différentes durées d'intégration T_{ild} (en s). Le signe de l'ILD (plus ou moins) indique la direction d'où provient la source sonore (gauche ou droite respectivement). Le signal, un jeu de clés agités de gauche à droite, est le même que Fig. 4.12. L'ILD est estimé avec un seuil τ fixé à 0 et une paire de 80 filtres gammatones disposés dans une bande entre 100 Hz et 8 kHz.

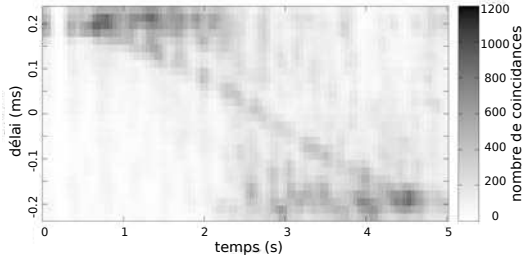
de la source de la gauche vers la droite. Néanmoins ce *pattern* est largement bruité. En plus d'un « bruit de fond » relativement constant, constitué de coïncidences détectées à des délais aberrants, s'ajoutent quelques zones extrêmement bruitées. À partir de 3.5 s environ, la majorité des coïncidences est ainsi détectée dans les délais positifs (autour de 0.2 ms) à l'opposé du délai théoriquement attendu, ce qui entraîne évidemment une mauvaise estimation de la direction de la source. Ces différents artefacts sont causés principalement par les conditions acoustiques (bruits et réverbération) mais également par les réflections multiples causées par les pavillons. Considérant maintenant la Fig. 4.15(b), où l'ITD est calculé avec un seuil $\tau = 10^{-2}$, l'estimation s'avère plus robuste que précédemment, la suppression des composantes de basse intensité ayant fortement réduit le bruit de fond constaté dans la configuration précédente. Il apparaît néanmoins que l'utilisation du seuil provoque une « binarisation » de l'estimation, les coïncidences étant majoritairement détectées soit complètement à gauche soit complètement à droite selon la direction de la source. Ce phénomène est bien visible vers 2.5 s lorsque la source passe de gauche à droite. Observons enfin la Fig. 4.15(c) où l'ITD est cette fois estimé sur les fronts d'ondes uniquement, avec un temps de demi-vie $\delta = 30$ ms. Le résultat obtenu est bien meilleur que dans les configurations précédentes et l'estimation s'avère très stable, confirmant ainsi l'intérêt de l'extraction des fronts d'ondes comme pré-traitement au calcul de l'ITD. Évidemment le nombre de coïncidences détectées est moindre que précédemment, le calcul s'effectuant sur les plus informatives d'entre elles.

4.3.2 Simulations de localisation

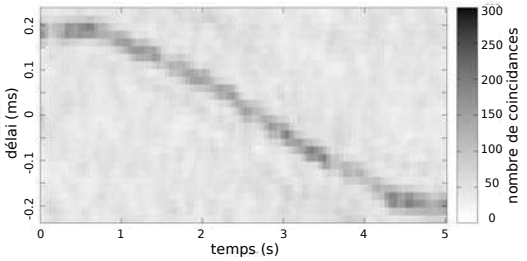
Maintenant que nous venons de tester notre modèle binaural sur un enregistrement correspondant à un environnement acoustique réaliste, illustrant ainsi de manière empirique ses capacités à l'audition spatiale, ce paragraphe revient sur certains points d'importance du modèle grâce à deux séries de simulations. La première vise à évaluer les capacités de localisation en milieu anéchoïque pour différentes sources, allant des tons purs à des spectres complexes, ce qui nous permettra également d'opérer



(a) $\tau = 0$, $\delta = 0$ ms.



(b) $\tau = 10^{-2}$, $\delta = 0$ ms.



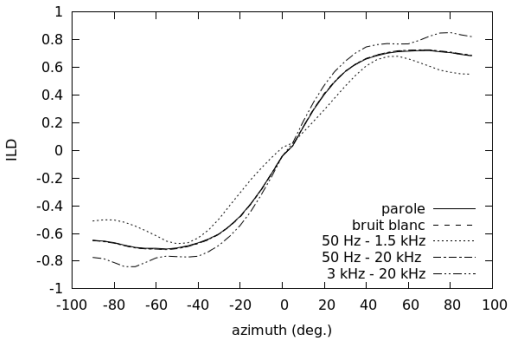
(c) $\tau = 0$, $\delta = 30$ ms.

FIGURE 4.15 – ITD obtenu à partir d'un enregistrement binaural sur Psikharpax. Le signal d'entrée est le même que Fig. 4.12. (a) ITD obtenu en conditions « neutres », avec l'extraction des fronts d'ondes désactivée (*i.e.* avec un temps de demi-vie $\delta = 0$ ms) et un seuil de transduction $\tau = 0$. (b) ITD estimé avec une demi-vie $\delta = 0$ ms et un seuil $\tau = 10^{-2}$. (c) ITD estimé avec une demi-vie $\delta = 30$ ms et un seuil $\tau = 0$. Le calcul est effectué sur une paire de 80 filtres gammatones disposés entre 100 Hz et 8 kHz.

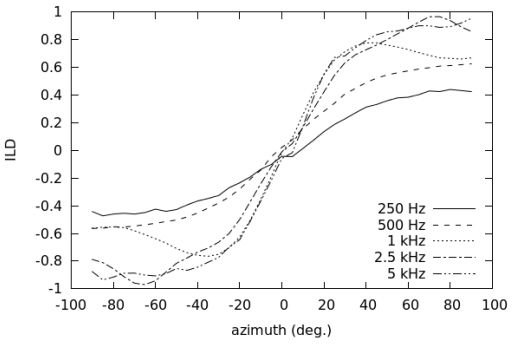
une courte digression autour de la théorie duplex et les domaines de pertinence des indices binauraux (voir paragraphe 2.1.1.2). La seconde série de simulations reviendra en détail sur l'extraction des fronts d'onde en milieu anéchoïque et réverbérant, et permettra encore une fois de démontrer la pertinence de cette méthode en guise de prétraitement à l'extraction de l'ITD.

4.3.2.1 Autour de la théorie duplex

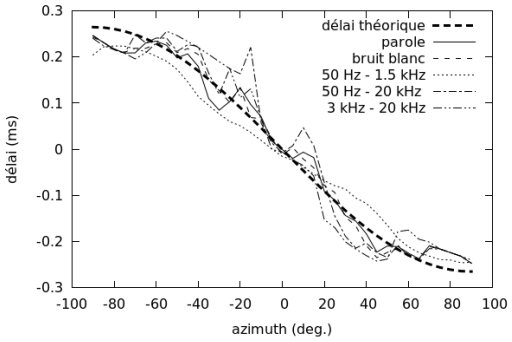
Cette première simulation évalue les indices d'ILD et d'ITD obtenus à partir de différentes sources sonores générées artificiellement. L'évaluation de ces différentes sources, couvrant des spectres basses ou hautes fréquences, nous permettra de caractériser les performances de localisation de notre modèle pour différents signaux, à la lumière de la théorie duplex prédisant un fonctionnement optimal de l'ILD dans les aigus et de l'ITD dans les graves (paragraphe 2.1.1.2). Un total de 10 sources est généré à une fréquence d'échantillonage $f_s = 44.1$ kHz : 5 tons purs, de simples sinus donc, dont la fréquence varie entre 250 Hz et 5 kHz, et 5 sources au spectre plus riche : un signal de parole (le mot « cochlée » dont le cochléogramme est affiché Fig. 4.6), un bruit blanc et trois sources constituées d'un mélange aléatoire de 600 sinusoïdes (voir l'Eq. 5.5), dans une bande de fréquence différente pour chacune des 3 sources : entre 50 Hz et 1.5 kHz, entre 3 kHz et 20 kHz et enfin entre 50 Hz et 20 kHz. Une fois les différentes sources générées, leur position dans l'espace et le filtrage de l'oreille externe sont simulés pour 36 azimuts différents, entre $-90°$ et $90°$ par pas de $5°$, en utilisant la méthode d'interpolation des HRTF présentée au paragraphe 4.1.1



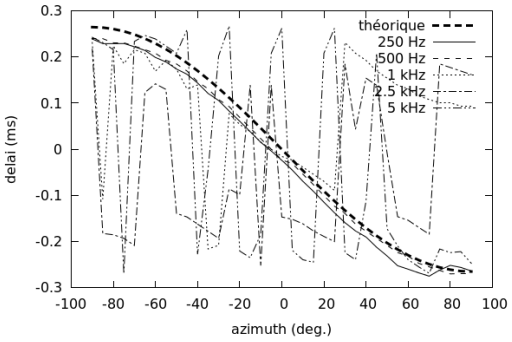
(a) ILD sur spectres larges.



(b) ILD sur tons purs.



(c) ITD sur spectres larges.



(d) ITD sur tons purs.

Figure 4.16 – Indices binauraux intégrés en fréquence pour différentes sources obtenues en simulation en conditions anéchoïques. (a)(b) ILD. (c)(d) ITD sans fronts d'onde. (a)(c) sources à large bandes incluant parole, bruit blanc et mélange de sinusoïdes pour 3 bandes fréquentielles. (b)(d) tons purs de 250 Hz à 5 kHz.

(modèle 3). Le filtrage cochléaire est opéré par un banc de 30 filtres gammatones situés entre 100 Hz et 8 kHz. Le seuil τ est nul et l'extraction des fronts d'onde est ici désactivée. Un temps d'intégration de 20 ms est retenu à la fois pour le calcul de l'ILD et de l'ITD, cette constante de temps étant suffisante pour le traitement des sources stationnaires utilisées ici. Cette valeur de 20 ms est de plus communément utilisée dans les modèles d'audition artificielle, notamment ceux basés sur une décomposition FFT.

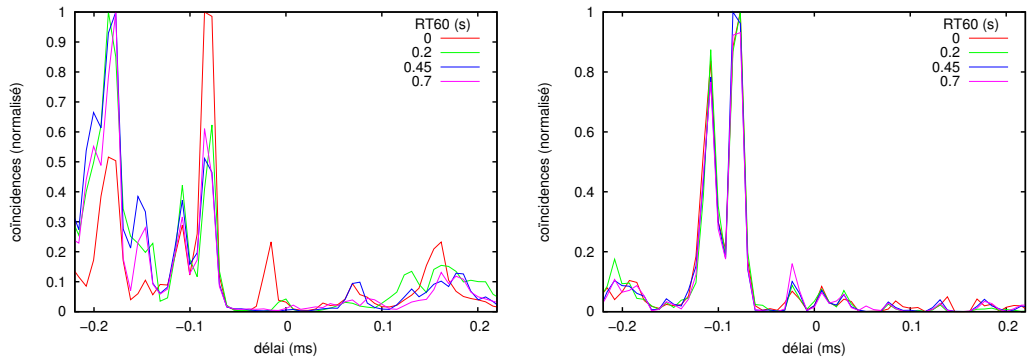
La Fig. 4.16 présente le résultat de cette simulation. Les indices binauraux sont intégrés en fréquence, c'est-à-dire que seule la moyenne des différents canaux cochléaires est considérée. Concernant l'ILD tout d'abord, nous observons que cet indice suit la même évolution quel que soit le type de source, la différence d'intensité étant tout de même moins marquée pour les sources basses-fréquences (50 Hz – 1.5 kHz, Fig. 4.16(a)) et particulièrement les tons purs (200 Hz et 500 Hz, Fig. 4.16(b)). Considérant maintenant les indices d'ITD des sources à spectre large (Fig. 4.16(c)), nous constatons que le délai estimé est proche du délai théorique (calculé conformément à l'Eq. 3.4) dans tous les cas. Observons tout de même que la source composée de hautes fréquences (3 kHz – 20 kHz) mène à une estimation correcte de l'ITD et contredit ainsi une vision « binaire » et littérale de la théorie duplex. Observons également l'estimation obtenue à partir de la source à basses fréquences. L'estimation du délai s'avère très stable et correspond vraisemblablement aux performances

maximales atteignables par notre modèle mais le délai estimé est systématiquement sous évalué par rapport au délai théoriquement attendu (en valeur absolue), ce qui est à rapprocher de la sous-estimation également rapportée chez l'humain sous le terme de flou de localisation (paragraphe 2.1.1.1). Concernant pour finir l'estimation de l'ITD sur des tons purs (Fig. 4.16(d)) nous observons que l'estimation du délai est aberrante au-delà de 500 Hz. Ce phénomène est attendu puisqu'il est la conséquence directe des ambiguïtés de phase inhérente à une longueur d'onde largement inférieure au diamètre de la tête, telle que prise en compte par la théorie duplex. Notons pour finir que l'exploitation des fronts d'ondes ne permet pas dans notre modèle d'améliorer ce résultat pour les tons purs en hautes fréquences.

4.3.2.2 Fronts d'ondes et réverbération

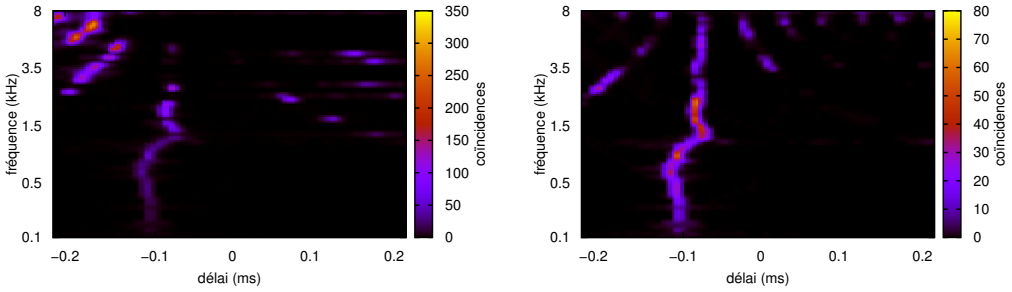
Nous avons vu au paragraphe 4.3.1 que l'extraction des fronts d'ondes permet d'accroître la précision et la robustesse de l'ITD, permettant ainsi de supprimer les coïncidences aberrantes associées aux impulsions secondaires. Cette simulation vise à évaluer l'intérêt des fronts d'ondes dans notre modèle en présence de réverbération. Les fronts d'ondes sont en effet généralement utilisés dans la littérature pour accroître la robustesse d'un système à ce type de perturbations (voir paragraphe 4.2.1.1). Dans cette expérience la réverbération est simulée par la génération de réponses impulsionnelles d'une chambre acoustique par la méthode image (Allen & Berkley, 1979) et nous utilisons pour ce faire l'algorithme proposé par Habets (2006). Nous simulons ainsi une pièce rectangulaire de dimensions 5m x 4m x 2.75m dans laquelle un auditeur, assimilé au point situé au centre de l'axe interaural, est placé à la position $[2, 2, 1.5]$ et une source sonore, de position constante, est placée par rapport à l'auditeur à une distance de 2 m pour un azimut de 30° et une élévation nulle. Différentes conditions acoustiques sont générées à travers la simulation de différents temps de réverbération à 60 dB (RT60), c'est-à-dire le temps mis par un signal acoustique pour voir son amplitude atténuée de 60 dB. Nous utilisons ainsi 4 RT60 différents : 0 ms (conditions anéchoïques), 200 ms, 450 ms et 700 ms. Ces valeurs sont utilisées par exemple par May *et al.* (2011) et Youssef *et al.* (2013) pour évaluer la robustesse de leur méthode de localisation en conditions réverbérantes. Une fois le signal source convolué par les réponses impulsionnelles ainsi générées, le filtrage des HRTF est appliqué comme précédemment par la méthode présentée au paragraphe 4.1.1 (modèle 3). Le modèle auditif utilisé ici se focalise sur l'ITD exclusivement. Nous utilisons un banc de 30 filtres cochléaires entre 100 Hz et 8 kHz, un seuil de transduction $\tau = 0$, un temps de demi-vie $\delta = 30$ ms lorsque l'extraction des fronts d'onde est activée et une durée d'intégration $T_{itd} = 100$ ms. La source sonore utilisée est un mélange de 100 sinusoïdes disposées entre 100 Hz et 8 kHz (Eq. 5.5).

La Fig. 4.17 illustre les résultats de cette simulation pour les 4 temps de réverbération simulés et avec l'extraction des fronts d'onde activée et désactivée. Les courbes d'ITD obtenues, représentant le nombre de coïncidences détectées pour chaque délai, sont normalisées. En effet nous sommes ici uniquement intéressés par le maximum d'ITD, associé à la position azimutale perçue. Observons tout d'abord que, en conditions anéchoïques (RT60 = 0 s, courbes rouges), le maximum d'ITD détecté autour de -0.1 ms et associé à un azimut de 30° est le même avec et sans fronts d'onde. L'utilisation des fronts d'onde permet néanmoins d'éliminer une grande partie des artefacts présents, à l'exception d'un pic secondaire. L'intérêt des fronts d'onde devient évident lorsque que l'on observe les résultats obtenus en conditions réverbérantes. Ainsi dans les 3 cas étudiés (200 ms, 450 ms et 700 ms), le maximum d'ITD obtenu sans fronts



(a) Fronts d'ondes désactivés (b) Fronts d'ondes activés

Figure 4.17 – Extraction de l'ITD avec et sans fronts d'onde en conditions réverbérantes. (a) Extraction des fronts d'ondes désactivée. (b) Extraction des fronts d'ondes activée. La source sonore est un mélange de 100 sinus compris entre 100 Hz et 8 kHz située à 2 m de distance et à un azimut de 30°. La courbe d'ITD représente le nombre de coïncidences détectées en fonction du délai sur toute la durée du signal. Le maximum de la courbe obtenue correspond à l'ITD dominant le signal et donc à l'azimut « perçu » par le modèle.



(a) Fronts d'ondes désactivés (b) Fronts d'ondes activés

Figure 4.18 – Détail de l'ITD obtenu en fonction de la fréquence pour RT60 = 0.45 s. (a) Extraction des fronts d'ondes désactivée. (b) Extraction des fronts d'ondes activée. Le protocole et les données sont les mêmes que Fig. 4.17.

d'onde est fortement corrompu par la présence de réverbération et est « dévié » vers les -0.2 ms : le pic à 0.1 ms est toujours présent mais n'est plus le maximum global de la courbe d'ITD. Remarquons de plus que les résultats obtenus dans les 3 conditions échoïques sont relativement similaires, la principale différence apparaissant ici en fonction de l'absence ou de la présence de réverbération. Nous constatons enfin que l'extraction des fronts d'ondes permet de rendre l'ITD insensible aux réverbérations, les courbes de la Fig. 4.17(b) étant similaires. Cependant dans les 4 cas étudiés avec fronts d'onde, la présence d'un maximum local secondaire est retrouvée autour de -0.12 ms.

La Fig. 4.18 présente le détail de l'ITD obtenu pour RT60 = 0.45 s en fonction du délai interaural et de la fréquence des canaux cochléaires, avec et sans extraction

des fronts d'ondes. Il est constaté que les artefacts perturbants l'estimation de l'ITD sont concentrés dans les hautes fréquences (Fig. 4.18(a)). Il est de plus montré que l'extraction des fronts d'ondes permet de réduire ces artefacts, rendant de la sorte l'estimation de l'ITD plus robuste aux conditions réverbérantes (Fig. 4.18(b)). Enfin le détail de la structure de l'ITD permet de déterminer que le pic secondaire est provoqué par un décalage du délai entre les hautes et basses fréquences, la limite se situant autour de 1.5 kHz.

4.4 Discussion

Le modèle auditif binaural présenté dans ce chapitre permet donc d'extraire les indices d'ILD et d'ITD nécessaires à la localisation binaurale. Cette discussion revient tout d'abord sur les aspects bioinspirés du modèle et la simplicité de certains choix de modélisation. Nous abordons ensuite les performances de ce modèle en termes de localisation. Enfin nous revenons sur le rôle des différents paramètres du modèle.

Choix de modélisation A chaque étape de traitement, nous avons motivé les choix de modélisation par des arguments biologiques, en tâchant de ne retenir de ce substrat anatomique et neuronal que les principes clés d'organisation. Le modèle présenté ici repose néanmoins sur des approximations très importante de la biologie. La représentation impulsionnelle décrite au paragraphe 4.1.3 illustre parfaitement ce choix de simplicité. En conservant les caractéristiques temporelles et d'intensité du signal d'entrée, ce modèle reproduit le trait fondamental de la transduction auditive par la simple détection de maxima locaux. Une modélisation plus fine de la transduction aurait ainsi exigé une base probabiliste, comme le modèle de CCI proposé par Sumner *et al.* (2002), mais aurait ajouté une complexité non nécessaire. Cette simplicité de modélisation se retrouve également dans le calcul des fronts d'onde par exemple, où le temps de réfraction suit une décroissance linéaire plutôt qu'exponentielle (la décroissance généralement utilisée dans l'état de l'art).

Partant du principe que le système auditif présente des imperfections dans la manière dont est captée puis codée l'information auditive (voir à ce sujet la discussion du chapitre 2), nous n'avons en effet pas cherché à concevoir un système binaural parfaitement calibré. Les pavillons du robot-rat Psikharpax permettent ainsi une amplification directionnelle fondamentale pour la localisation, mais les HRTF produites n'ont pas été quantifiées et sont dans le détail certainement très éloignées de HRTF humaines ou animales. Cette mesure ne s'est pas avérée nécessaire puisque, dans le cadre de la robotique autonome, un modèle ne peut prendre en compte ce type d'information *a priori*. De plus la méthode de localisation sensorimotrice proposée aux chapitres suivants permet d'apprendre ces HRTF de manière implicite. De plus, les microphones sont fixés aux pavillons par de la colle, méthodologie approximative qui induit vraisemblablement des erreurs de calibration (les microphones ne sont pas exactement orientés dans le même axe) mais permet en retour de créer une asymétrie des capteurs pouvant être exploitée pour la résolution de l'ambiguïté avant-arrière (voir paragraphe 6.1.2).

Nous proposons ainsi un modèle binaural optimisé pour la robotique, privilégiant l'efficacité algorithmique à la complexité des calculs mis en oeuvre. Ce modèle sera utilisé dans la suite de cette thèse dans un contexte sensorimoteur. En ce sens sa simplicité est un atout puisqu'elle nous permettra de nous focaliser sur les aspects essentiels des interactions sensorimotrices. Par ailleurs, les différentes approxima-

tions ne sont pas incompatibles avec l'approche sensorimotrice. Selon cette théorie, nos capacités perceptives émergent en effet des interactions sensorimotrices et sont indépendantes du type de codage utilisé (voir paragraphe 5.1.2).

Localisation de source sonore Concernant la partie expérimentale présentée au paragraphe 4.3, les capacités du modèle à la localisation d'une source unique, y compris en conditions acoustiques difficiles, est ici établie grâce à l'étude de cas concrets. Cette capacité sera la base des chapitres suivants dans lesquels nous nous intéresserons à la localisation active et à l'apprentissage de l'espace auditif sans revenir sur l'évaluation des indices binauraux. Néanmoins une étude plus poussée concernant le modèle dans son ensemble permettrait d'établir avec précision le comportement des indices binauraux dans un cadre multisource par exemple, et de mieux appréhender l'effet des paramètres efférents dans ces conditions. De plus une comparaison avec des modèles classiques de l'état de l'art, à base de FFT et de GCC notamment, serait une valeur ajoutée importante à notre travail. Nous pensons ici particulièrement à notre implémentation du modèle de Jeffress, dont les expériences précédentes ont démontré la robustesse dans des conditions où les modèles basés corrélation ont été analysés comme peu stables, en conditions échoïques notamment.

Paramètres du modèle La simplicité du modèle binaural évoquée ci-dessus est également due au faible nombre de ses paramètres. Si l'on oublie les paramètres statiques, fixé à l'initialisation du modèle (nombre de canaux ou fréquence d'échantillonage par exemple), ces paramètres sont au nombre de 4 (seuil de transduction, demi-vie des fronts d'ondes, durée d'intégration en ITD et en ILD, Fig. 4.8 et 4.11). Ils permettent d'adapter la réponse du modèle aux changements des conditions environnementales (bruit, réverbération).

Dans la partie expérimentale de ce chapitre, nous avons fixé ces valeurs à la main soit en les faisant varier sur une large amplitude dans le but d'évaluer leur influence sur les indices binauraux obtenus, soit en se focalisant sur une valeur unique mettant en exergue un phénomène particulier. Cette manière de procéder n'est bien sur pas convenable dans le contexte de la robotique autonome puisqu'en fixant une valeur, nous insérons un *a priori* que nous espérons adapté au problème posé.

La discussion générale de cette thèse revient sur cette modélisation du système efférent. Il semble en effet que, sous certaines hypothèses, la théorie sensorimotrice puisse offrir une manière originale de traiter ce problème. Nous avons vu en effet au chapitre 3 que les systèmes efférents sont en général sous-exploités dans les modèles auditifs artificiels, illustrant donc le bénéfice que l'on pourrait tirer d'une telle modélisation, notamment dans un contexte multisource où les méthodes actuelles trouvent leurs limitations.

Chapitre 5

Approche sensorimotrice de la localisation

Sommaire

Ce chapitre applique la théorie des contingences sensorimotrices au problème de la localisation de sources sonores. Cette théorie stipule que la perception est par nature un phénomène actif, et place donc l'action au coeur du processus de perception. Dans ce contexte, la localisation de source est assimilée à une interaction sensorimotrice qu'un agent entretient avec son environnement, sans qu'aucune connaissance *a priori* de cet environnement ne soit nécessaire. Nous avons vu de plus aux chapitres 2 et 3 qu'il existe une différence entre d'une part la nature fondamentalement active du processus de localisation de sources sonores en biologie et d'autre part l'approche majoritairement passive des modèles artificiels. Ainsi la théorie sensorimotrice présente un double intérêt dans le cadre de la localisation de sources sonores

pour la robotique autonome. Elle permet d'adresser d'une part le problème de la dépendance du processus de localisation à certains paramètres extérieurs (connaissance des HRTF, *a priori* sur le bruit, le nombre de source, leur spectre, etc), mais elle permet également d'inscrire la localisation dans un processus actif fondé sur une interaction de l'agent avec son environnement.

La première partie de ce chapitre est consacrée à la présentation de la théorie sensorimotrice et des principaux arguments justifiant cette approche. Les quelques applications proposées par la littérature sont ensuite introduites, dans le domaine de la perception de l'espace et celui plus spécifique de la localisation de sources sonores. La seconde partie du chapitre présente alors une formalisation sensorimotrice de la localisation de source sonore, puis introduit une première méthode de localisation. Cette méthode, fonctionnant sur une base passive et supervisée, n'a pas vocation à être utilisée dans le contexte de la robotique autonome. Elle servira néanmoins de base aux chapitres suivants qui se focaliseront sur les aspects actifs et non-supervisés de la tâche de localisation. Enfin la section expérimentale présente une évaluation de cette première méthode pour la localisation supervisée, à la fois en azimut et en élévation. Ces expériences seront également l'occasion de nous intéresser à deux problèmes classiques de la littérature concernant la localisation de sources sonores : ceux de l'ambiguïté avant-arrière et de la localisation en élévation.

5.1 Approche sensorimotrice de la perception

Ce premier paragraphe présente tout d'abord quelques résultats qui remettent en cause l'approche « classique » de la perception et motivent la théorie des contingences sensorimotrices. Cette théorie est ensuite introduite dans le cadre de la perception de l'espace. Enfin un état de l'art des quelques applications de cette théorie est présenté.

5.1.1 Limitations de l'approche « classique »

Nous avons vu au chapitre 3, et particulièrement au paragraphe 3.1.1, que les modèles de systèmes auditifs se composent généralement d'étapes successives de traitements de plus en plus abstraits. Ce schéma est commun à la quasi-totalité des approches suivies en robotique, et illustre en fait une organisation interne de type *sentir-planifier-agir* héritant directement des origines de la recherche en intelligence artificielle (Laflaquière, 2013). Cette conception de la perception que nous qualifions de « classique » introduit, nous l'avons vu, de nombreux problèmes : manque d'adaptabilité et difficulté de modélisation d'environnements complexes notamment. De nombreux arguments expérimentaux viennent mettre en difficulté cette vision classique de la perception, ceux-ci sont répertoriés par O'Regan (2011). Nous présentons ici trois expériences en psychologie de la perception qui remettent en cause le schéma *sentir-planifier-agir*.

Ainsi Held & Hein (1963) propose une expérience cherchant à évaluer le rôle des mouvements volontaires de chatons sur leur perception visuelle. Deux chatons sont placés à l'ouverture de leur paupières dans un environnement visuel formé de bandes verticales noires et blanches. Le premier chaton est placé dans un chariot et ne peut pas se déplacer librement. Le second chaton est libre de ses mouvements et entraîne avec lui le chariot du premier, le système étant conçu afin de garantir une expérience visuelle identique au deux chatons. Après quelques jours dans cet environnement, les chatons sont confrontés à des tâches visuelles simples. Si les chatons restés libres

de leurs mouvements accomplissent ces tâches avec succès, les chatons restés passifs échouent. Cette expérience suggère ainsi que l'action est une composante nécessaire à l'acquisition de capacités visuelles.

Un autre argument est celui de la substitution sensorielle (Bach-y Rita *et al.* , 1969; Bach-y Rita & Kercel, 2003; Auvray *et al.* , 2005). Dans leurs travaux, Bach-y Rita *et al.* (1969) proposent un mécanisme de substitution de la modalité visuelle par la modalité tactile chez des sujets aveugles. Une caméra placée sur une paire de lunettes retranscrit l'image capturée sous forme d'impulsions mécaniques provoquées par une matrice de « picots » placée sur le ventre ou sur le dos du sujet. Après une phase d'apprentissage, l'identification et la spatialisation d'objets est possible et l'expérience tactile se substitue alors à l'expérience visuelle, les deux expériences se révélant similaires (O'Regan & Noë, 2001). Néanmoins cette substitution ne s'opère que si la caméra est activement manipulable, signe là encore du rôle fondamental des mouvements volontaires dans le processus de perception.

Enfin, le dernier argument que nous présenterons ici est lié à la cécité au changement (Noë & O'Regan, 2000; Auvray & O'Regan, 2003). En effet des expériences en psychologie de la vision ont démontré que des changements très importants apportés à une scène visuelle peuvent passer inaperçu à la plupart des observateurs. Ces changements peuvent être brusques (deux personnes inversent leurs chapeaux, un immeuble grossit de 25%, la couleur d'un perroquet passe subitement du rouge au vert, etc) mais doivent intervenir de façon détournée, soit pendant une saccade oculaire, soit être dissimulés par un distracteur. Des changements progressifs peuvent également passer inaperçu, par exemple une couleur d'arrière plan passant graduellement de l'orange au vert. Cette dernière expérience recoupe donc le constat formulé au paragraphe 2.4 au sujet des « imperfections » des systèmes auditif et visuel, à savoir que nos systèmes perceptifs ne procèdent pas à une reconstruction parfaite et fidèle de notre environnement.

5.1.2 Théorie des contingences sensorimotrices

La théorie des contingences sensorimotrices prend sa source dans les intuitions de Poincaré (1895) concernant les fondements de la géométrie et la perception de l'espace. La position de Poincaré est que notre perception de l'espace n'est pas une faculté innée mais qu'elle se construit par le biais de notre expérience sensorimotrice, c'est-à-dire des retours perceptifs que nous avons de nos propres actions. L'idée générale est que le système nerveux central peut s'assimiler initialement à un agent naïf qui communique avec le monde extérieur à travers un ensemble inconnu de connexions afférentes et efférentes, c'est-à-dire de voies sensorielles et de voies motrices. En considérant de plus cet agent comme n'ayant aucune connaissance *a priori* sur l'espace dans lequel il est immergé, l'approche sensorimotrice (Poincaré, 1895; O'Regan & Noë, 2001; Philipona *et al.* , 2003; Frolov, 2011) suggère que le cerveau analyse les conséquences des mouvements qu'il commande sur les perceptions qu'il reçoit en retour. Cette analyse donne ainsi accès à la structure et aux propriétés de l'espace physique, aux propriétés du corps dans lequel il est incarné et, plus généralement, à toutes les notions sensorielles auxquelles notre cerveau peut se rattacher (Frolov, 2011).

A la vue de ces différents arguments, et sur la base des intuitions de Poincaré (1895), O'Regan & Noë (2001) proposent ainsi que la vision, et plus généralement la perception, se défini non pas comme une succession de processus passifs augmentant à chaque étape le degré d'abstraction mais, à l'inverse, par la capacité à modifier ac-

tivement ses sensations, par le mouvement ou l'attention, grâce à la connaissance acquise par l'expérience sensorimotrice des transformations liant actions et sensations. Les capacités perceptives sont dans cette approche indépendantes des capteurs et du type de traitements opérés, comme l'illustre par exemple le phénomène de la substitution sensorielle, par laquelle on parvient à « voir avec le toucher ». L'hypothèse de base sur laquelle repose cette approche sensorimotrice est que l'espace perceptif d'un agent, potentiellement de très haute dimension (puisque l'on peut considérer initialement chaque fibre nerveuse comme une dimension indépendante), est inclus dans une variété de basse dimension dont la topologie est homéomorphe à l'espace physique extérieur (Philipona *et al.* , 2003). Par voie de conséquence, l'apprentissage de la perception de l'espace se concrétise par l'apprentissage d'une telle variété sensorielle. Pour parvenir à cet apprentissage, le système nerveux se base sur la compensabilité des changements sensoriels qu'il perçoit et plus précisément sur la détection de *mouvements compensables* : un changement sensoriel induit par l'environnement peut être compensé par un déplacement de l'agent, ou *vice versa*, l'agent retrouvant alors sa sensation initiale. Selon l'approche sensorimotrice, c'est donc cette propriété de compensabilité de nos changements sensoriels qui est à la base de notre perception de l'espace. Ce type de traitements est vraisemblablement à l'oeuvre dans le système nerveux central (Doya, 1999; Frolov, 2011) puisque le cervelet a été identifié pour son rôle dans la prédiction sensorielle des mouvements volontaires (Philipona *et al.* , 2003). O'Regan (2011) va bien plus loin que la simple perception spatiale et aborde notamment le problème de la conscience, mais ces considérations d'ordre philosophique dépassent le cadre de cette thèse.

5.1.3 Applications en perception active

Ce paragraphe présente les quelques applications de la théorie sensorimotrice proposées dans la littérature. Nous introduisons en premier lieu les études se rapportant à la perception de l'espace en général, puis plus spécifiquement les travaux consacrés à la localisation de sources sonores.

5.1.3.1 Perception de l'espace

Nous l'avons dit au paragraphe précédent, l'approche sensorimotrice considère un agent totalement naïf et dépourvu d'*a priori* sur son environnement ou son propre corps. Du point de vue robotique, cela implique que le système ne dispose pas de modèle géométrique de l'espace ou du robot, de même il n'a pas accès aux concepts de position ou d'objet. Philipona *et al.* (2003) démontrent qu'un tel agent, à partir des relations entre entrées perceptives et sorties motrices, est en mesure premièrement de distinguer ses connexions extéroceptives de ses connexions proprioceptives et deuxièmement, sur la base de cette connaissance et de l'analyse des transformations compensables, d'en déduire la dimensionnalité de l'« espace physique » dans lequel l'agent évolue. Pour cela, les auteurs considèrent l'analyse du flux sensorimoteur dans trois conditions différentes à l'aide de trois simulations : (1) des changements environnementaux interviennent sans changement de la configuration motrice, (2) des changements moteurs interviennent dans un environnement stationnaire et (3) des changements interviennent à la fois dans l'environnement et dans la configuration motrice de l'agent. Ces travaux valident donc l'approche sensorimotrice dans son application à la perception de l'espace.

À partir d'une simulation de robot équipé de capteurs visuels et auditifs, Couverture & Gas (2009) ont ainsi proposé une implémentation de la méthode de Philipona *et al.* (2003) et sont parvenus à retrouver la dimension de l'espace physique d'un agent naïf. Les résultats obtenus par les auteurs ne restent néanmoins valides que pour des mouvements infinitésimaux (rotations de l'ordre de 10^{-5} ou 10^{-6} degrés), rendant cette approche inapplicable dans un contexte robotique réaliste. Ainsi Laflaquière *et al.* (2010) ont proposé une approche en partie non-linéaire au problème de l'estimation de dimension, permettant cette fois d'atteindre des amplitudes de mouvements raisonnables dans un contexte robotique (de l'ordre du degré). Roschin *et al.* (2011) proposent également une méthode alternative à l'apprentissage de la perception tridimensionnelle. Dans cette simulation, un bras articulé est fixé à un corps équipé de capteurs tactiles. L'apprentissage, effectué par un réseau de neurones, est ici basé sur la coïncidence des sensations tactiles et proprioceptives lors d'un auto-contact, c'est-à-dire lorsque le bout du bras touche le propre corps de l'agent.

L'apprentissage de la dimensionnalité de l'espace n'est cependant que la première étape et l'approche sensorimotrice peut également être appliquée à l'apprentissage de concepts plus complexes. Laflaquière (2013) propose ainsi un ensemble de simulations dans lesquelles un agent naïf est amené à découvrir des notions de l'espace de plus en plus complexes grâce à différentes stratégies d'exploration. Il est ainsi montré qu'un tel agent est en mesure de se construire une représentation de l'espace indépendante de l'environnement. La notion de transformation compensable permet alors la découverte d'une régularité dans l'espace sensorimoteur, dont dérivent les notions de translation et de distance.

En plus de ces travaux présentant des résultats en simulation, quelques applications de la théorie sensorimotrice ont été proposées en robotique. Au-delà de ces quelques travaux présentés ci-dessous, quelques travaux en robotique tentent également de dépasser les limitations imposées par l'approche classique, en s'inspirant parfois implicitement de la théorie sensorimotrice (Laflaquière, 2013). Nous pouvons ainsi citer les domaines de la robotique développementale (Lungarella *et al.* , 2003), la robotique évolutionniste (Harvey *et al.* , 2005) ou encore la robotique référencée capteur (Chaumette & Hutchinson, 2006).

5.1.3.2 Localisation de sources sonores

Dans une approche appliquée plus spécifiquement à la localisation de sources sonores, Aytekin *et al.* (2008) ont démontré qu'un agent naïf peut apprendre à localiser des sources sonores, sans aucun *a priori* sur ses HRTF ni aucune forme d'expérience en terme d'audition spatiale, uniquement sur la base des conséquences sensorielles de mouvements volontaires. Le seul *a priori* requis est la connaissance de la dimensionnalité de l'espace auditif, qui est dans ce cas bidimensionnel (variation des sources en azimut et en élévation). Les auteurs ont ainsi proposé un modèle d'apprentissage de l'espace auditif basé sur la théorie sensorimotrice et ont démontré ses capacités en simulation. L'espace auditif est appris sur des HRTF humaines et de chauve-souris provenant de différents sujets et, après une décomposition FFT du signal, c'est la différence d'intensité avant et après un faible mouvement en rotation (1° seulement) qui est prise en compte pour différentes positions de la source. Sur la base de ces vecteurs tangents, un algorithme non-linéaire d'estimation de variétés basé sur l'alignement des espaces tangents locaux (LTSA) (Zhang & Zha, 2002) est appliqué. Les variétés obtenues respectent globalement la topologie de l'espace

auditif extérieur. Les auteurs remarquent cependant que l'apprentissage s'effectue difficilement dans certaines zones, notamment pour une élévation importante, où la variété obtenue présente localement une distribution non-uniforme associée à une forte courbure. Notons néanmoins que la question de la localisation de nouvelles sources après apprentissage n'est pas adressée par Aytekin *et al.* (2008), ce qui réduit d'autant l'intérêt pratique de cette approche. Les auteurs discutent néanmoins de la pertinence de leur modèle pour la robotique, où l'ensemble des contraintes ou paramètres à prendre en compte par un modèle peuvent difficilement l'être *a priori*.

Deleforge & Horaud (2011) étendent l'approche proposée par Aytekin *et al.* (2008) vers la robotique en proposant une expérimentation basée sur des enregistrements effectués sur la tête binaurale présentée Fig. 3.1(c). Ces enregistrements effectués en conditions acoustiques réalistes incluent 16200 positions à différents angles d'azimut (à 360°) et d'élévation (à 180°). Différents indices liés à l'intensité du signal sont alors calculés, à partir desquels des représentations en basse dimension sont apprises par l'algorithme LTSA. Les auteurs montrent que les représentations obtenues sont homéomorphes à la variété motrice d'une tête à deux degrés de liberté en rotation, c'est-à-dire que chaque point sur cette représentation peut être associé à une coordonnée motrice de la tête binaurale. Via la loi cinématique de cette tête, les auteurs remontent alors à la position de la source dans l'espace physique.

Enfin quelques méthodes robotiques indirectement basées sur l'approche sensorimotrice ont été proposées pour la localisation de sources sonores et sont décrites au chapitre 3. Ces méthodes utilisent notamment une régression linéaire supervisée (Hörnstein *et al.* , 2006) et des cartes auto-organisatrices (Berglund *et al.* , 2008).

5.2 Formalisation

La théorie sensorimotrice a été formalisée mathématiquement par Philipona *et al.* (2003). Après avoir introduit cette formalisation, ce paragraphe propose une définition sensorimotrice de la localisation puis propose une méthode de localisation reposant sur un contexte passif et supervisé. Cette première méthode sera par la suite enrichie aux chapitres 6, 7 et 8, toujours sur la base du formalisme proposé ici.

5.2.1 Espace sensoriel, espace moteur et loi sensorimotrice

La formalisation proposée ici pour la théorie sensorimotrice se base, nous l'avons dit, sur les travaux de Philipona *et al.* (2003, 2004). Ce paragraphe présente successivement les notions d'espace sensoriel, d'espace moteur et de loi sensorimotrice, cette dernière permettant d'exprimer les sensations expérimentées par un agent en fonction de l'environnement et de son propre état moteur.

Espace sensoriel Soit s le vecteur représentant les entrées sensorielles d'un agent à un instant donné. Il associe une dimension indépendante à chaque capteur (chaque pixel d'une image ou chaque fréquence d'un spectrogramme constituent une dimension) et, avec n_s le nombre de dimensions, nous avons $s \in \mathbb{R}^{n_s}$. Les valeurs de s sont cependant contraintes, notamment par la disposition ou la fonction de transfert des capteurs, ou encore la variabilité de l'environnement. Considérons ainsi $\mathcal{S}$ comme l'ensemble des sensations expérimentables par l'agent, que nous appelons *espace sensoriel*. Du fait des contraintes imposées par les capteurs, $\mathcal{S}$ constitue donc un sous-ensemble de $\mathbb{R}^{n_s}$.

Après Philipona *et al.* (2003), nous faisons l'hypothèse que l'espace sensoriel $\mathcal{S}$ est une variété plongée dans $\mathbb{R}^{n_s}$. Une variété est un espace topologique qui peut être localement assimilée à un espace euclidien (il existe une relation bijective continue entre tout voisinage de $\mathcal{S}$ et cet espace euclidien). Intuitivement cette variété est semblable a une surface ayant subi une déformation non-rigide quelconque (courbure, étirement) et plongée dans un espace de plus grande dimension (ici $\mathbb{R}^{n_s}$).

Espace moteur Notons également $m \in \mathcal{M}$ le vecteur représentant les sorties motrices émises par l'agent à un moment donné, $\mathcal{M}$ étant l'ensemble des états moteurs accessibles à l'agent ou *espace moteur*. Nous supposons là encore que $\mathcal{M}$ est une variété plongée dans $\mathbb{R}^{n_m}$, où n_m est la dimension de l'espace moteur. Chaque dimension correspond ici a un degré de liberté de l'agent. Sous l'hypothèse que cet agent ne possède aucune connaissance *a priori*, ni sur l'environnement qui l'entoure ni sur sa propre structure corporelle, s et m sont les seules informations par lesquelles l'agent peut interagir avec le monde extérieur.

Loi sensorimotrice Poursuivons notre raisonnement et notons maintenant $p \in \mathcal{P}$ la configuration du corps de l'agent. $\mathcal{P}$ correspond à l'ensemble des états corporels que peut prendre l'agent. Ainsi p représente les positions et orientations des membres et articulations de l'agent dans l'espace extérieur. Suivant toujours Philipona *et al.* (2003), l'état corporel p de l'agent est contrôlé par l'état moteur m au travers de la loi cinématique Φ_a de l'agent. Notons également $e \in \mathcal{E}$ la configuration de l'environnement à un instant donné, $\mathcal{E}$ est l'*espace environnemental* correspondant à l'ensemble des états environnementaux expérimentables par l'agent. e représente ainsi tous les paramètres permettant de définir un environnement. Dans le cadre de la localisation de sources sonores, e peut inclure la position et l'orientation d'une source, son spectre ou encore des données sur la réverbération. Les entrées sensorielles s sont alors déterminées en fonction de p et de e selon une fonction Φ_b. Nous avons ainsi :

$$p = \Phi_a(m) \text{ et } s = \Phi_b(p, e). \tag{5.1}$$

Bien sur l'expression des fonctions Φ_a et Φ_b de même que les valeurs de p et de e sont inaccessibles à l'agent, s et m sont en effet les seules informations dont dispose l'agent sur ses interactions avec le monde extérieur. Nous définissons finalement la fonction Φ composant les 2 fonctions Φ_a et Φ_b et reliant donc directement l'entrée sensorielle s à la sortie motrice m et à l'état de l'environnement e. La loi Φ, que nous appelons *loi sensorimotrice* ou encore « loi sensorimotrice fonctionnelle » selon Philipona, s'exprime alors comme :

$$s = \Phi(m, e) = \Phi_b(\Phi_a(m), e). \tag{5.2}$$

5.2.2 Définition sensorimotrice de la localisation

Nous avons vu au chapitre 3 que la majorité des méthodes de localisation cherchent à estimer un angle ou une position dans un repère donné *a priori*, et supposent ainsi la préexistence de la notion d'espace. La théorie sensorimotrice impose à l'inverse de redéfinir la localisation non plus en termes spatiaux mais directement en termes moteurs. Nous proposons ici une définition mathématique et purement motrice de la localisation de source. Cette définition est ici limitée au cas le plus simple d'une source unique en l'absence de bruit. L'extension de cette approche à des environnements plus complexes est néanmoins abordée dans la discussion générale de cette

thèse. Considérons ainsi une variété motrice $\mathcal{M}$ et un état environnemental $e \in \mathcal{E}$, nous définissons la localisation d'une source sonore comme l'estimation de l'état moteur $\tilde{m}$ tel que :

$$\tilde{m} = \underset{m \in \mathcal{M}}{\operatorname{argmin}} |\Phi(m, e) - \Phi(m_{ref}, e_{ref})|, \tag{5.3}$$

où $|.|$ représente une métrique de distance dans $\mathcal{S}$. La configuration $s_{ref} = (m_{ref}, e_{ref})$ représente l'état sensoriel de référence que l'estimation de $\tilde{m}$ cherche à atteindre. Cette sensation s_{ref} est initialement inconnue de l'agent, elle doit donc être soit apprise, soit donnée *a priori*. Dans le cadre de la localisation nous définissons s_{ref} comme la sensation générée par une source localisée en face de l'agent (état environnemental e_{ref}) dont la tête est placée en position de repos (état moteur m_{ref}). Cette configuration correspond au cas le plus évident de localisation. Le chapitre 6 propose de plus un modèle de localisation active permettant d'estimer cet état s_{ref} grâce à un comportement d'orientation de la tête vers la direction de la source.

L'existence de cet état sensoriel s_{ref} que l'agent cherche à atteindre par l'estimation de l'état moteur $\tilde{m}$ est donc l'hypothèse centrale de notre approche. Cette définition rentre de plus directement dans le cadre de pensée de Poincaré (1895) qui, lorsque qu'il parle de la localisation, écrit [1] : « Quand on dit que nous *localisons* tel objet en tel point de l'espace, qu'est-ce que cela veut dire ? Cela signifie simplement que nous nous représentons les mouvements qu'il faut faire pour atteindre cet objet ; et qu'on ne dise pas que pour se représenter ces mouvements, il faut les projeter eux-mêmes dans l'espace et que la notion d'espace doit, par conséquent, préexister. Quand je dis que nous nous représentons ces mouvements, je veux simplement dire que nous nous représentons les sensations musculaires qui les accompagnent et qui n'ont aucun caractère géométrique, qui par conséquent n'impliquent nullement la préexistence de la notion d'espace. » Poincaré rejette donc toute nécessité de connaître l'espace physique pour parvenir à la localisation, seule intervenant les représentations motrices et sensorielles.

Un tel encodage de commandes motrices relatives plutôt qu'absolues est de plus retrouvé dans le monde vivant. Krauzlis *et al.* (1997) (voir également Oertel *et al.* (2001)) démontrent en effet la présence d'une carte motrice dans SC exploitée pour les saccades oculaires dans laquelle les neurones ne codent pas un état moteur absolu mais plutôt une « erreur motrice » codant le mouvement que les yeux doivent effectuer pour positionner au centre du champ visuel un stimulus auditif et/ou visuel.

5.2.3 Localisation passive et supervisée

Ce paragraphe propose une méthode de localisation qui se base sur un échantillonnage de l'espace sensorimoteur pour interpoler une estimation de $\tilde{m}$. Elle est supervisée puisque l'échantillonnage sensorimoteur est donné *a priori*. Elle est de plus passive puisqu'elle ne nécessite pas le déplacement de l'agent. Bien que cette méthode ne soit ainsi pas adaptée à la robotique autonome, celle-ci sera à la base des chapitres suivants qui proposeront des modèles de localisation active.

5.2.3.1 Échantillonnage de l'espace sensorimoteur

L'estimation de $\tilde{m}$, qui procède par une minimisation dans l'espace sensorimoteur, nécessite la connaissance des variétés $\mathcal{S}$ et $\mathcal{M}$. On doit de plus connaître la relation qui lie ces deux variétés dans le cadre de la localisation, c'est-à-dire que chaque point

1. H. Poincaré. La science et l'hypothèse. Flammarion, Champs sciences, édition 2009, p. 82.

de $\mathcal{S}$ doit pouvoir être associé au point de $\mathcal{M}$ estimant la position de la source en termes moteurs.

Or du point de vue de l'agent il est impossible d'accéder directement à l'expression analytique de ces variétés ou à la relation sensorimotrice, seul est possible un échantillonnage de celles-ci. L'échantillonnage de l'espace sensorimoteur fourni à l'agent est ainsi composé de 3 éléments : un échantillonnage S de l'espace sensoriel $\mathcal{S}$, un échantillonnage M de l'espace moteur $\mathcal{M}$, tous deux de taille n, et une carte sensorimotrice A reliant S et M point à point. Nous avons ainsi $S = \{s_i\}_{i\in[1,n]}$, $M = \{m_i\}_{i\in[1,n]}$ et $A = \{(s_i, m_i)\}_{i\in[1,n]}$. Cette association est donnée dans ce chapitre de manière supervisée (paragraphe 5.3), mais elle sera établie de manière active aux chapitres 6 et 7.

5.2.3.2 Interpolation dans l'espace sensorimoteur

Supposons un agent ayant à sa disposition ces 3 ensembles S, M et A formant un échantillonnage de son espace sensorimoteur. Il s'agit maintenant de pouvoir localiser de nouvelles source à partir de cet échantillonnage. Ce paragraphe propose ainsi une méthode d'estimation de $\tilde{m}$ basée sur une interpolation dans l'échantillonnage sensorimoteur. Cette estimation se base sur une classification aux plus proches voisins (Fix & Hodges, 1951; Stone, 1977; Wang *et al.* , 2007; Peterson, 2009). Cette méthode fut initialement proposée pour les problèmes de classification ou de régression pour lesquels peu ou pas d'*a priori* sont disponibles sur la distribution d'entrée. En considérant l'existence des échantillonnages sensoriel et moteur S et M respectivement, cette méthode repose sur la seule hypothèse que 2 points voisins dans S sont associés à 2 points voisins dans M. Plus formellement ceci revient donc à supposer que la fonction Φ est continue autour du voisinage considéré.

Considérant donc une nouvelle sensation $\tilde{s} \in \mathcal{S}$ associée à un environnement e constant, l'estimation $\tilde{m} \in \mathcal{M}$ de son état moteur, comme définit par l'Eq. 5.3, s'effectue à partir des relations de voisinage de $\tilde{s}$ dans S. Appelons $K_{\tilde{s}} = \{s_i | s_i \in S\}_{i\in[1,k]}$ l'ensemble des k plus proches voisins (k-ppv) de $\tilde{s}$ dans S et $K_{\tilde{m}} = \{m_i | m_i \in M\}_{i\in[1,k]}$ l'ensemble des états moteurs correspondants aux éléments de $K_{\tilde{s}}$ dans A. L'estimation $\tilde{m}$ de l'état moteur correspondant à $\tilde{s}$ est finalement calculée à partir des ensembles $K_{\tilde{s}}$ et $K_{\tilde{m}}$ par interpolation, en utilisant une pondération inverse à la distance (Shepard, 1968). Nous avons ainsi :

$$\tilde{m} = \sum_{i=1}^{k} \frac{w_i m_i}{\sum_{j=1}^{k} w_j}, \text{ avec } w_i = \frac{1}{|\tilde{s} - s_i|}. \tag{5.4}$$

Cette pondération inverse à la distance garantit une plus faible influence des points les plus éloignés de $\tilde{s}$ et, à l'inverse, une contribution plus importante de ses voisins les plus proches.

La suite de cette thèse ne fait pas de différence entre l'espace physique et l'espace moteur : les valeurs d'un état moteur $m \in \mathcal{M}$ refléteront directement un angle en degrés dans l'espace physique. Cette simplification permettra d'alléger les équations mais n'est pas en soit une limitation à la méthode proposée. En effet l'agent n'ayant pas accès à la géométrie du monde extérieur, n'importe quelle transformation de cet espace pourrait être utilisée de manière totalement transparente, sous la contrainte tout de même que l'échantillonnage autour de m soit suffisamment dense pour capturer la topologie locale de la variété motrice.

L'annexe C présente de plus l'opportunité de procéder à une interpolation aux k-ppv non pas directement dans S mais par l'intermédiaire d'une représentation en basse dimension de S. Ce type de traitements est utilisé notamment par Aytekin *et al.* (2008), Deleforge & Horaud (2011), Laflaquière (2013) au sein de l'approche sensorimotrice, des méthodes de réduction de dimension étant en effet utilisée pour représenter la topologie de l'espace sensoriel avec un nombre moindre de paramètres. L'annexe C discute de l'intérêt d'un tel prétraitement à l'interpolation et montre qu'il n'apporte pas d'amélioration des performances de localisation. La réduction de dimension sera cependant utilisée au paragraphe 5.3.2 à des fins de visualisation de l'espace sensoriel.

5.3 Expériences

Cette section présente deux expériences principales. La première propose une mise en oeuvre de la méthode proposée ci-dessus dans une tâche de localisation en azimut et en élévation. La seconde expérience se focalise sur le rôle des HRTF dans la localisation azimutale. Il est ainsi montré que la présence d'indices spectraux dans les données sensorielles permet de supprimer l'ambiguïté avant/arrière.

5.3.1 Localisation passive et supervisée

Le propos de cette expérience est de valider la méthode de localisation par interpolation proposée au paragraphe 5.2.3.2. Cette expérience utilise des signaux binauraux provenant d'une base de donnée dédiée à la localisation binaurale, l'échantillonnage de l'espace sensorimoteur est donc fourni de manière supervisée. Ce paragraphe détaille tout d'abord le protocole expérimental, puis présente les résultats obtenus en termes de localisation. Contrairement aux idées répandues dans la littérature, il est montré que la localisation en élévation est possible à partir des indices binauraux classiquement utilisés pour la localisation en azimut.

5.3.1.1 Protocole expérimental

Ce paragraphe présente tout d'abord la base de donnée d'enregistrements binauraux utilisée pour cette série d'expérience. Le calcul des indices binauraux et le processus de localisation sont ensuite détaillés.

La base de donnée CAMIL La série d'expériences que nous proposons ici repose sur la base de donnée CAMIL proposée par Deleforge & Horaud (2011). Cette base de donnée se compose d'enregistrements binauraux effectués sur la plateforme robotique illustrée Fig. 3.1(c) dans un environnement de bureau, donc en présence de réverbération et d'un léger bruit de fond. Une source sonore, placée à une distance fixe de 2.7 m, diffuse un bruit stationnaire composé de 600 sinusoïdes aléatoires durant 1.1 s. Les différentes fréquences composant le signal sont fixes tandis que les phase et amplitude sont aléatoirement tirées sur une loi uniforme. Plus précisément, un tel signal $x(t)$ s'exprime en fonction du temps comme (Deleforge & Horaud, 2011) :

$$x(t) = \frac{\sum_{i=1}^{n} \omega_i \sin(2\pi f_i t + \phi_i)}{\sum_{i=1}^{n} \omega_i}, \tag{5.5}$$

Table 5.1 – Paramètres utilisés dans les expériences du paragraphe 5.3.

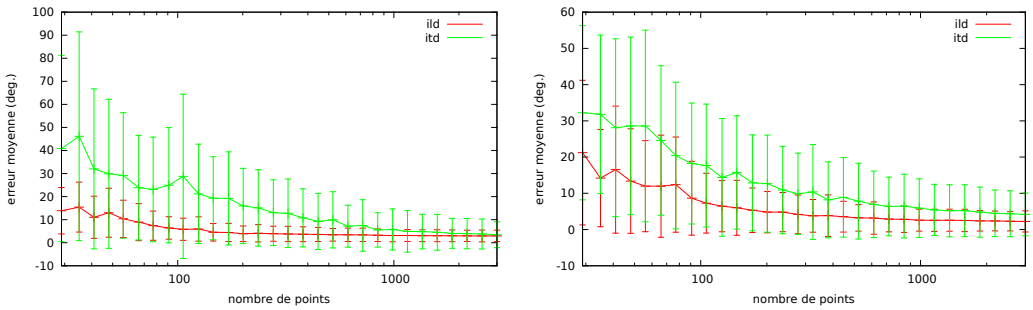
Cochlée		
n	nombre de canaux cochléaires	30
f_{min}	fréquence de résonance minimale (Hz)	100
f_{max}	fréquence de résonance maximale (Hz)	8000
τ	seuil de transduction	0
Indices binauraux		
d_{inter}	distance interaurale (m)	0.19
T_{ild}	intégration de l'ILD (ms)	100
T_{itd}	intégration de l'ITD (ms)	100
δ	demi-vie du seuil de fronts d'onde (ms)	3
Interpolation		
k	ordre du voisinage	12

où $F = \{f_i\}_{i=1..n}$ est un ensemble de n fréquences fixées, $\{\omega_i\}_{i=1..n} \in [0,1]^n$ et $\{\phi_i\}_{i=1..n} \in [0,2\pi]^n$ étant respectivement les poids et phases associés à chaque fréquence. Un jeu de $n = 600$ fréquences dans l'ensemble $F = [50, 150, 250, ..., 5950]$ Hz est utilisé. L'échantillonnage, fixé initialement à 48 kHz, est sous-échantillonné à 20 kHz dans le but de réduire le temps de calcul des indices binauraux (voir l'annexe D.1 pour des précisions sur l'exécution du modèle binaural). Les enregistrements sont effectués pour un azimut compris dans l'intervalle $[-180°, 180°]$ et une élévation comprise dans l'intervalle $[-90°, 90°]$, ces deux angles variant par pas de $2°$, pour un total de 16200 positions différentes. À chaque enregistrement est associée la vérité terrain, permettant ainsi un apprentissage supervisé de l'espace auditif. Nous utilisons de plus un signal de référence pour l'apprentissage et des signaux aléatoires pour les tests, tel que proposé par Deleforge & Horaud (2011).

Concernant les changements de positions de la source notons toutefois que, pour des raisons techniques, ce n'est pas la position de la source qui varie mais celle de la tête binaurale. Ainsi les changements d'états environnementaux liés aux changements de position de la source - changements dans la structure de la réverbération ou encore du bruit de fond - ne sont pas reflétés dans cette base de donnée, ce qui réduit la complexité du problème et facilite d'autant le processus d'apprentissage.

Indices binauraux À partir de ces sources sonores sont générés un ensemble de vecteurs d'ILD et d'ITD au travers du modèle binaural présenté au chapitre 4, les paramètres utilisés étant résumés Table 5.1. Pour chaque position, un total de 5 vecteurs sont générés pour chaque indice (ILD et ITD), nous donnant donc un total de 81000 vecteurs, indépendamment pour les ensembles d'apprentissage et de test.

Les vecteurs d'ILD sont de dimension 30, correspondant au nombre de canaux cochléaires, tandis que les vecteurs d'ITD sont de dimension 690, correspondant au nombre de canaux multipliés par le nombre de délais associés à une fréquence d'échantillonnage à 20 kHz et une distance interaurale de 19 cm (23 délais différents, voir l'Eq. 4.12). Notons à ce titre que, la distance interaurale réelle du système n'étant pas fournie par Deleforge & Horaud (2011), nous avons établi cette valeur à la main, en considérant les indices d'ITD obtenus pour un azimut maximal (*i.e.* à $\pm 90°$).



(a) en azimut (b) en élévation

FIGURE 5.1 – Évolution de l'erreur d'interpolation moyenne $\Delta\tilde{m}$ pour 2000 vecteurs de test selon la taille de l'espace d'apprentissage (de 30 à 3000 points sur une échelle logarithmique), en ILD (rouge) et en ITD (vert). (a) erreur moyenne en azimut, erreur finale en ILD à $2.9 \pm 2.5°$, en ITD à $3.5 \pm 5.6°$. (b) erreur moyenne en élévation, erreur finale en ILD à $2.2 \pm 2.9°$, en ITD à $4.2 \pm 5.9°$. Les barres d'erreur représentent l'écart-type.

Localisation Considérant maintenant un état sensoriel $\tilde{s}$ de l'ensemble de test et un ensemble d'apprentissage composé de n points et représenté par une carte sensorimotrice $A = \{(s_i, m_i)\}_{i\in[1,n]}$ telle qu'introduite au paragraphe 5.2.3.2, l'interpolation de l'état moteur $\tilde{m}$ associé à $\tilde{s}$ s'effectue par interpolation inverse à la distance dans A, comme proposé par l'Eq. 5.4. Puisque nous nous situons ici dans un espace moteur bidimensionnel (azimut et élévation), avec $m = (m_\theta, m_\phi)$ l'état moteur correspondant à la vérité terrain associée à $\tilde{s}$, l'erreur d'interpolation $\Delta\tilde{m}$ associée à l'estimation $\tilde{m}$ est définie comme :

$$\Delta\tilde{m} = \sqrt{(m_\theta - h(\tilde{m}_\theta))^2 + (m_\phi - \tilde{m}_\phi)^2}, \tag{5.6}$$

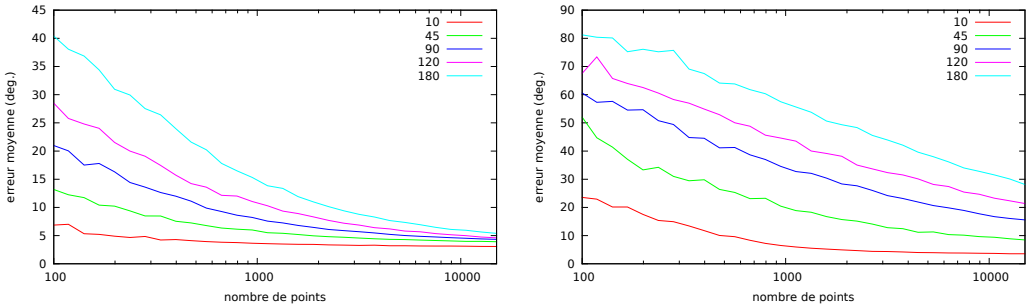
où la fonction $h(\tilde{m}_\theta)$ permet de prendre en compte la périodicité de l'axe azimutal et consiste simplement à ramener l'erreur d'estimation $\tilde{m}_\theta$ dans l'intervalle $[-180°, 180°]$. Nous avons ainsi :

$$h(\tilde{m}_\theta) = \begin{cases} \tilde{m}_\theta - 360 & \text{si } \tilde{m}_\theta > 180, \\ \tilde{m}_\theta + 360 & \text{si } \tilde{m}_\theta < 180, \\ \tilde{m}_\theta & \text{sinon.} \end{cases} \tag{5.7}$$

5.3.1.2 En fonction de la taille de l'échantillonnage

Cette première expérience présente une analyse de l'erreur de localisation en fonction de la taille de l'échantillonnage. Nous considérons trois scénarios de localisation distincts : localisation en azimut seulement ($m_\theta \in [-180°, 180°]$ et $m_\phi = 0$), localisation en élévation seulement ($m_\theta = 0$ et $m_\phi \in [-90°, 90°]$) et enfin localisation conjointe en azimut et en élévation ($m_\theta \in [-180°, 180°]$ et $m_\phi \in [-90°, 90°]$).

En azimut La Fig. 5.1(a) présente l'évolution de l'erreur de localisation $\Delta\tilde{m}$ moyenne obtenue en azimut. L'erreur est donnée en fonction de la taille de l'ensemble d'apprentissage, variant de 30 à 3000 points, pour un ensemble de test constant composé de 2000 points. Les résultats sont présentés en ILD et en ITD. Il est premièrement constaté que l'erreur obtenue décroît en fonction du nombre de points, de



(a) Évolution de l'erreur moyenne en ILD (b) Évolution de l'erreur moyenne en ITD

Figure 5.2 – Évolution de l'erreur de localisation conjointe en azimut et en élévation en fonction de la taille n de l'espace d'apprentissage (variant de 100 à 15000 selon une échelle logarithmique) et de l'intervalle d'élévation pris en compte. La taille de l'ensemble de test est fixée à 5000 points. (a) erreur moyenne en ILD. (b) erreur moyenne en ITD. Les sources ont un azimut dans l'intervalle $\pm 180°$ et un intervalle d'élévation de $\pm 5°$ (rouge), $\pm 22.5°$ (vert), $\pm 45°$ (bleu), $\pm 60°$ (violet) et $\pm 90°$ (cyan).

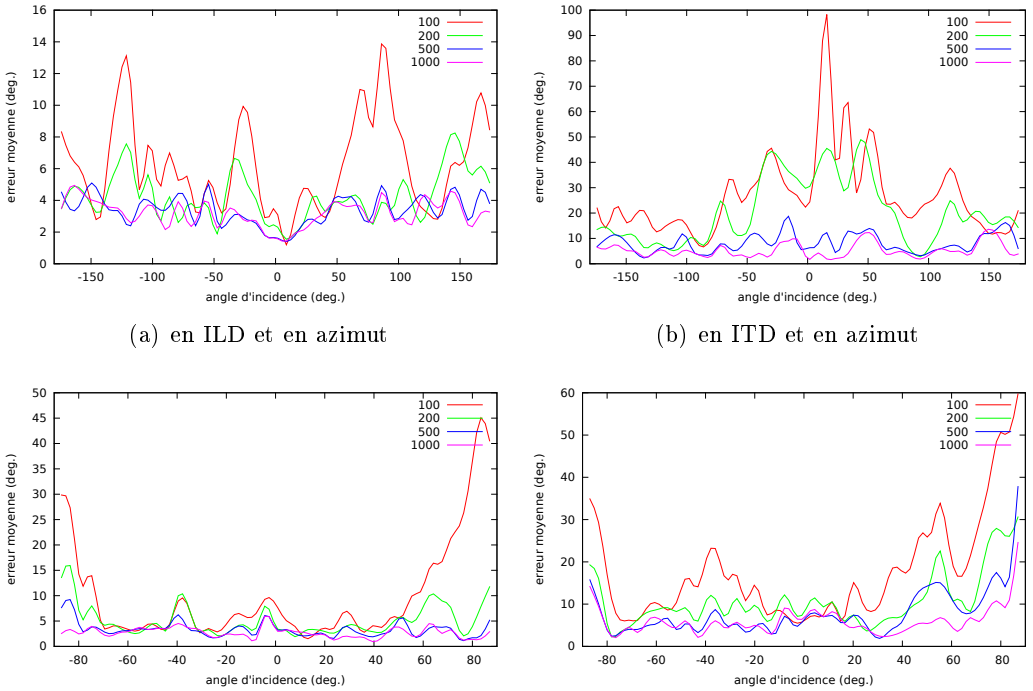
même que l'écart type associé. Ce résultat est attendu puisque un ensemble d'apprentissage plus dense signifie un voisinage plus proche et donc une interpolation plus précise. Enfin, l'indice d'ITD présente des performances moindre que celles obtenues avec l'ILD, ce qui s'explique par sa plus grande dimensionnalité. Asymptotiquement néanmoins les 2 indices semblent converger vers des performances similaires. Ainsi pour un ensemble d'apprentissage composé de 3000 points, l'erreur moyenne en azimut est de 2.9° en ILD contre 3.5° en ITD

En élévation La même expérience reproduite en élévation est présentée Fig. 5.1(b). Ces données appellent les mêmes constats que pour le cas azimutal : les performances progressent avec la densité de l'ensemble d'apprentissage et les erreurs en ILD et en ITD semblent converger asymptotiquement. L'erreur moyenne obtenue pour un ensemble d'apprentissage composé de 3000 points est de 2.2° en ILD contre 4.2° en ITD.

En azimut et en élévation La Fig. 5.2 résume les données obtenues pour une localisation conjointe en azimut et en élévation. L'erreur est donnée en fonction d'intervalles d'élévation de plus en plus larges, variant de $\pm 5°$ à $\pm 90°$.

Concernant l'ILD tout d'abord (Fig. 5.2(a)), si l'erreur de localisation associée à une faible élévation est similaire à celle obtenue Fig. 5.1, cette erreur augmente significativement avec l'élargissement de l'intervalle d'élévation lorsque l'ensemble d'apprentissage est réduit. Néanmoins avec la croissance de l'ensemble d'apprentissage, les erreurs associées aux différents intervalles convergent vers une erreur moyenne de 5° environ.

Concernant maintenant l'indice d'ITD (Fig. 5.2(b)), les mêmes conclusions peuvent être tirées mais avec, comme précédemment, des performances moindre qu'en ILD. Une interpolation correcte exige en effet en ITD un nombre très important de points comparé à l'ILD, là encore du fait de la différence de dimensionnalité. Ainsi la convergence de l'erreur n'est toujours pas observée par interpolation sur un ensemble d'apprentissage de 15000 points, la taille maximale retenue ici.



(a) en ILD et en azimut

(b) en ITD et en azimut

(c) en ILD et en élévation

(d) en ITD et en élévation

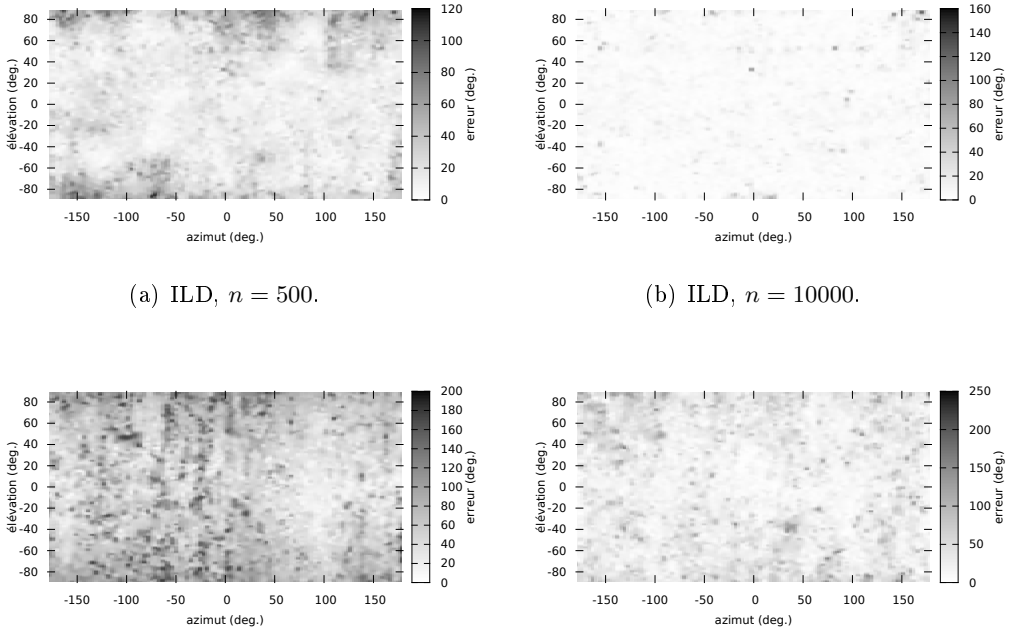
FIGURE 5.3 – Évolution de l'erreur moyenne $\Delta\tilde{m}$ en fonction de la position de la source en azimut (a, b) et en élévation (c, d) pour l'ILD (a, c) et l'ITD (b, d). La taille de l'ensemble de test est fixée à 2000 points, celle de l'ensemble d'apprentissage varie entre 100 (rouge), 200 (vert), 500 (bleu) et 1000 points (violet).

5.3.1.3 En fonction de la direction de la source

Si l'expérience précédente analyse l'erreur moyennée sur toutes les directions, cette seconde expérience détaille les performances de localisation obtenues en fonction de la direction de la source. Nous considérons là encore trois scénarios de distincts : localisation en azimut seulement, en élévation seulement et enfin localisation conjointe en azimut et en élévation.

En azimut La Fig. 5.3(a) détaille la distribution de l'erreur en fonction de l'azimut de la source pour différentes tailles de l'ensemble d'apprentissage (variant de 100 à 1000 points). L'erreur associée à l'ILD est relativement homogène sur l'ensemble de l'axe azimutale, avec une meilleure performance autour de 0°. À l'inverse, l'erreur associée à l'ITD (Fig. 5.3(b)) est concentrée dans un intervalle centré en face de l'agent tandis que la localisation de sources provenant de son dos s'effectue avec une erreur plus faible.

En élévation Concernant maintenant la localisation en élévation (Fig. 5.3(c) et 5.3(d)), il est montré que, tant en ILD qu'en ITD, l'erreur associée à un ensemble d'apprentissage réduit est concentrée dans les directions périphériques. Cet effet peut s'expliquer par le fait que ces directions se trouvent sur le « bord » de l'espace moteur, et donc que le voisinage associé à l'interpolation s'étend dans une seule direction, entraînant un biais et une sous-évaluation systématique de $\tilde{m}$.



(a) ILD, $n = 500$.

(b) ILD, $n = 10000$.

(c) ITD, $n = 500$.

(d) ITD, $n = 10000$.

Figure 5.4 – Répartition de l'erreur moyenne en ILD et en ITD pour $n = 500$ ou $n = 10000$ échantillons dans l'espace d'apprentissage. (a) ILD, $n = 500$. Erreur moyenne à $18 \pm 16°$. (b) ILD, $n = 10000$. Erreur moyenne à $4.1 \pm 4.9°$. (c) ITD, $n = 500$. Erreur moyenne $= 65 \pm 38°$. (d) ITD, $n = 10000$. Erreur moyenne $= 19 \pm 21°$.

En azimut et en élévation La Fig. 5.4 présente finalement l'erreur de localisation obtenue en fonction de l'azimut et de l'élévation de la source, en ILD et en ITD. L'erreur est calculée pour deux ensembles d'apprentissage composés respectivement de 500 et de 10000 points. Pour 500 points, l'erreur en ILD est essentiellement concentrée dans les élévations extrêmes tandis qu'elle est répartie de manière plus homogène en ITD. Lorsque l'ensemble d'apprentissage passe à 10000 points nous constatons en ILD comme en ITD que, si une large erreur est associée à des positions précises, de large zones de l'espace auditif permettent une interpolation dans de bonnes conditions.

5.3.2 Apprentissage de l'ambiguïté avant/arrière

L'expérience précédente a montré que la méthode de localisation est en mesure de résoudre l'ambiguïté avant/arrière. Si une erreur de localisation est localement présente sur l'axe azimutal, il n'y a en effet pas de confusion entre les directions avant et arrière (Fig. 5.3(a) et 5.3(b)). Les chapitres 2 et 3 ont montré qu'il s'agit là d'une capacité de localisation non-triviale, l'humain utilisant par exemple de légers mouvements de tête pour lever cette ambiguïté.

Cette expérience, dont les résultats sont tirés en partie de Bernard *et al.* (2012), approfondit ce point en analysant l'espace sensoriel obtenu à partir de différents

modèles d'oreille externe. Il est ainsi montré que ce sont les indices spectraux fournis par les HRTF qui permettent de lever l'ambiguïté avant/arrière.

5.3.2.1 Protocole expérimental

A l'inverse de l'expérience précédente qui présente des résultats de localisation, cette expérience se focalise sur l'analyse de l'espace sensoriel sans se préoccuper de l'aspect moteur. Ce paragraphe détaille la génération de 3 échantillonnages, chacun associé à une HRTF différente puis introduit une méthode de réduction de dimension permettant l'analyse et la visualisation des échantillonnages sensoriels.

Génération des échantillonnages L'agent est supposé immobile et l'environnement est composé d'une source sonore unique diffusant un spectre large-bande (Eq. 5.5). La source est placée aléatoirement à un azimut dans l'intervalle $[-180, 180]$ pour une élévation nulle et une distance de 2.7 m. Le modèle binaural utilise les paramètres de la Table 5.1 et seuls les indices d'ILD sont considérés.

Trois espaces sensoriels sont générés sur cette base, chacun associé à une HRTF différente : (1) un filtrage purement directionnel, sans indices spectraux (paragraphe 4.1.1.4), (2) une HRTF mesurée sur une tête binaurale incluant des indices spectraux (paragraphe 4.1.1.3 et (3) les enregistrements de la base CAMIL (Deleforge & Horaud, 2011) contenant bruit et réverbération en plus des indices spectraux. Ces trois espaces sont échantillonnés par 2000 états sensoriels, chacun associée à une source différente dont l'azimut est tiré aléatoirement selon une loi uniforme.

Les deux premiers échantillonnages sont générés avec un temps d'intégration fixé à $T_{ild} = 10^{-2}$ s contre $T_{ild} = 10^{-1}$ s pour l'échantillonnage à base d'enregistrements. L'intégration plus longue permet dans le second cas d'augmenter la robustesse des indices.

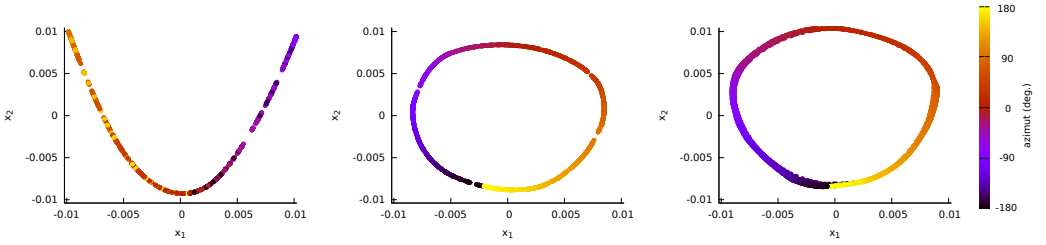
Réduction de dimension Les échantillonnages obtenus sont ensuite traités par l'algorithme des cartes propres Laplaciennes (LE), un algorithme de réduction de dimension qui nous permet ici de visualiser la topologie de chaque espace sensoriel dans une dimension réduite. Le principe de la réduction de dimension et l'algorithme du LE sont détaillés dans l'annexe C.

De la dimension initiale égale à 30 (le nombre de canaux cochléaires), les données sont réduites en dimension 2. Puisque seul l'angle azimutal de la source différencie les différentes sensations de l'agent, la dimension intrinsèque de la variété sensorielle est égale à 1. Cependant puisque cette variété contient un cycle, une dimension supplémentaire est nécessaire pour ne pas briser sa topologie.

5.3.2.2 Résultats

Les variétés apprises par la méthode des LE à partir des vecteurs ILD dans ces 3 configurations sont illustrées Fig. 5.5. Ces résultats montrent que les variétés permettent de représenter la direction gauche-droite d'une source sonore pour les trois configurations proposées. Cette expérience montre de plus que les indices liés à l'intensité du signal ne suffisent pas pour l'estimation de la direction avant-arrière (Fig. 5.5(a)) et que les indices spectraux apportés par les HRTF sont requis pour lever l'ambiguïté (Fig. 5.5(b).

Cette expérience a donc montré que la résolution avant-arrière peut se faire à partir des indices spectraux fournis par les HRTF, un simple filtrage périphérique



(a) Simulation (h^L_{dir}, h^R_{dir}). (b) Simulation (h^L_{hrtf}, h^R_{hrtf}). (c) Enregistrements CAMIL.

Figure 5.5 – L'apprentissage de l'ambiguïté avant/arrière dépend de la présence d'indices spectraux dans les HRTF. 3 échantillonnages sensoriels de 2000 points réduits à 2D par la méthode des LE. Chaque espace est associé à une HRTF différente. Les données sont obtenues en ILD, avec un azimut dans l'intervalle $[-180, 180]$ (dégradé de couleurs). (a) Sans indices spectraux, les directions gauche et droite sont discriminées mais les directions avant et arrière sont confondues. (b) L'ajout d'indices spectraux lève complètement l'ambiguïté avant/arrière. (c) Ceci reste vérifié pour les données calculées à partir de la base CAMIL.

purement directionnel n'étant pas suffisant pour parvenir à cette désambiguïté. En conséquence cette expérience montre également que l'espace sensoriel contient de manière implicite la structure des HRTF.

5.4 Discussion

Nous revenons tout d'abord sur l'application de la théorie sensorimotrice à la localisation de sources sonores puis nous évoquons l'aspect passif et supervisé du modèle et son extension à des capacités actives qui sera proposé aux chapitres suivants. Enfin nous évoquons plus spécifiquement la méthode de localisation et les performances obtenues.

Localisation sensorimotrice Le formalisme de la théorie des contingences sensorimotrices nous a permi de proposer une définition purement sensorimotrice de la tâche de localisation. Dans le cadre de la robotique autonome, le principal intérêt de cette approche est de s'affranchir de tout *a priori* sur l'environnement et le robot (à l'exception de la distance interaurale qui doit être connue pour l'extraction de l'ITD).

La méthode proposée dans ce chapitre permet une localisation passive (sans déplacement) grâce à une interpolation dans l'échantillonnage sensorimoteur. Cependant cet échantillonnage est fourni à l'agent de manière supervisée. Le chapitre 6 propose ainsi un comportement de localisation active permettant à l'agent de s'orienter en direction d'une source sonore. Ce comportement est alors utilisé au chapitre 7 pour échantillonner l'espace sensorimoteur de manière active et proposer une méthode d'apprentissage de la localisation qui soit totalement autonome, c'est-à-dire indépendante de toute connaissance autre que sensorielle et motrice.

Notons enfin que, selon la définition proposée par l'Eq. 5.3, la localisation est limitée à une source unique. En effet le formalisme introduit dans ce chapitre pour la localisation se situe au plus bas niveau sensorimoteur, n'a pas vocation à la localisation de sources multiples. Le chapitre 2 a en effet montré que la perception de sources

multiples fait appel à des processus moteurs et cognitifs complexes résumés par l'effet « cocktail party ». Nous reviendrons sur ce point dans la discussion générale de cette thèse.

Interpolation aux k voisins La méthode de localisation par interpolation que nous appliquons au paragraphe 5.3 repose sur une régression aux k-ppv. Elle offre des résultats satisfaisants en termes de localisation en azimut et, de manière plus surprenante, en élévation, qu'elle soit basée sur les indices d'ILD ou d'ITD. Néanmoins le cas de la localisation conjointe en ITD (Fig. 5.2(b)) fait apparaître une limitation bien connue de cette approche (et d'une manière générale à toute méthode basée sur un calcul de distance) : la dimensionnalité trop importante de l'espace d'entrée nécessite un échantillonnage dense et donc un nombre très élevé d'échantillons. Cette méthode, parmi les plus simples proposées dans le domaine de l'apprentissage artificiel, est à la base de nombre de méthodes plus complexes étudiant les caractéristiques locales de la distribution en entrée, comme l'apprentissage de variétés évoqué au paragraphe 5.3.2.1. Il est de plus prouvé que l'erreur de classification aux k-ppv converge vers le double de l'erreur de Bayes (l'erreur minimale théorique) lorsque $k = 1$ et $n \to +\infty$ (Peterson, 2009).

Dans le contexte de l'apprentissage autonome, cette approche présente l'avantage de ne reposer sur aucun *a priori* concernant la distribution d'entrée, hormis l'existence d'une métrique sur celle-ci. Puisque la distribution d'entrée est ici une variété sensorielle considérée localement, nous assimilons sa topologie à celle d'un espace euclidien (*i.e.* nous assumons un échantillonnage local de la variété suffisamment dense) et nous utilisons la distance euclidienne. Le choix de $k = 12$ a été choisi arbitrairement et bien que des méthodes d'estimation du k optimal aient été proposées, notamment à partir de validation croisée (Peterson, 2009), cet aspect n'a pas été approfondi dans cette thèse. L'inconvénient majeur de cette méthode est sa lourdeur computationnelle, du fait de l'utilisation massive de calculs de distances. Là encore des améliorations de cette méthode existent qui permettent de calculer une approximation des k-ppv, basées notamment sur des arbres kd ou des diagrammes de Voronoï. Ces optimisations n'ont néanmoins pas été abordées dans ce travail.

Performances de localisation Le paragraphe 5.3.1 a démontré la capacité de la méthode de localisation à obtenir de bons résultats sur des sources sonores réalistes. Si les performances en azimut sont attendues de part l'utilisation de l'ITD et de l'ILD, il est plus surprenant de constater les bons résultats obtenus en élévation. En effet il est communément admis que la localisation en élévation ne s'effectue non pas à partir d'indices binauraux mais grâce à des indices spectraux monauraux reflétant le filtrage des HRTF (voir le paragraphe 2.1.1.2). Cependant nous avons montré au paragraphe 5.3.2 que ces indices liés aux HRTF sont implicitement présents dans les indices binauraux et permettent l'apprentissage passif de l'ambiguïté avant-arrière. Ce même phénomène se produit ici, le filtrage des HRTF modifie la structure des indices binauraux, permettant leur utilisation pour la localisation en élévation.

Alors que la majorité des modèles binauraux exploités en robotique se concentrent sur les indices d'ITD, qui permettent d'estimer directement l'angle azimutal de la source, nous constatons que les performances de localisation sont meilleures en ILD qu'en ITD. Cette observation est également rapportée par Youssef *et al.* (2012b), qui ont montré que la capacité à différencier des azimuts est meilleure sur la base d'ILD qu'à partir d'ITD.

Chapitre 6

Localisation par orientation et déplacement

Sommaire

Le chapitre précédent propose une méthode de localisation passive (sans déplacement) basée sur une interpolation dans l'espace sensorimoteur. Cette méthode suppose cependant l'existence d'un échantillonnage de l'espace sensorimoteur, sur lequel est basée l'interpolation. Une série d'expérience a démontré l'intérêt de cette méthode pour la localisation de sources sonores. La connaissance de l'espace sensorimoteur était cependant apportée par le contexte supervisé et l'utilisation d'une base de donnée.

Ce chapitre propose une méthode de localisation alternative, dans laquelle un comportement réflexe se substitue à la connaissance *a priori* de l'espace sensorimoteur. Un réflexe met en oeuvre un cycle court de perception-action. La brièveté de ce cycle est à mettre en opposition à des processus plus complexes impliquant une cognition de haut niveau (par exemple l'utilisation d'une représentation de l'espace sensorimoteur). Dans le comportement réflexe, la perception est ainsi directement reliée à une réponse motrice, sans qu'une représentation interne ou une connaissance des interactions sensorimotrices ne soit impliquée. Cela en fait un cas élémentaire de perception active.

Différents comportements réflexes participent au processus de localisation auditive chez les mammifères (paragraphe 2.1.3), certains ont également fait l'objet d'une modélisation et d'une implémentation dans un contexte robotique (paragraphe 3.2.3.3). Ces solutions s'inspirent le plus souvent directement de la biologie, que ce

soit d'un point de vue psychologique avec les véhicules de Braitenberg (1986) notamment, ou d'un point de vue neurocomputationnel avec la modélisation du comportement d'orientation chez la chouette effraie (Rucci *et al.* , 2000) ou de la phonotaxie chez le grillon (Reeve & Webb, 2003).

Ce chapitre présente successivement deux comportements réflexes permettant la localisation active d'une source sonore. Chacun est introduit de manière formelle puis validée dans la section expérimentale. Le premier réflexe est un comportement d'orientation qui permet à l'agent d'orienter sa tête dans la direction azimutale d'une source sonore. Une capacité de déplacement est ensuite ajoutée au comportement d'orientation, permettant un mouvement en translation de l'agent vers de l'origine de la source. Ces deux processus actifs se basent sur le système binaural décrit au chapitre 4, et plus spécifiquement sur l'indice d'ILD.

6.1 Formalisation

Ce paragraphe présente les deux comportements réflexes proposés pour la localisation active de sources sonores. Le cadre sensorimoteur introduit au paragraphe 5.2 est utilisé.

6.1.1 Comportement d'orientation

Le comportement d'orientation permet la rotation du cou dans la direction azimutale d'une source sonore grâce à la minimisation de l'ILD. La position de la source est alors donnée *a posteriori*, c'est-à-dire une fois la minimisation effectuée.

6.1.1.1 Minimisation de l'ILD

Considérons un espace environnemental $\mathcal{E}$ composé d'une source sonore unique et stationnaire, disposée dans un environnement anéchoïque. L'azimut de la source e_ϕ de la source peut varier dans l'intervalle $[-90°, 90°]$, $0°$ correspondant à la direction directement en face de l'agent. $\mathcal{E}$ est donc de dimension 1, correspondant à la direction azimutale. Considérons de plus un espace moteur $\mathcal{M}$ également de dimension 1, permettant une rotation de l'axe du cou en azimut. Le comportement d'orientation se base sur l'indice d'ILD intégré en fréquence, de sorte que, en considérant le vecteur d'ILD $s^{ild}(t)$ obtenu par l'Eq. 4.14 à l'instant t, le vecteur sensation $s(t) \in \mathcal{S}$ s'exprime comme la moyenne de l'ILD sur les fréquences de $s^{ild}(t)$:

$$s(t) = \frac{1}{n} \sum_{i=1}^{n} s_i^{ild}(t), \tag{6.1}$$

où n correspond au nombre de canaux fréquentiels. L'espace sensoriel $\mathcal{S}$ est donc inclut dans $\mathbb{R}$ et correspond à l'intervalle $[-1, 1]$.

Le comportement d'orientation permet de minimiser la valeur absolue de $s(t)$ grâce à une rotation constante de l'axe du cou. Dans le but d'initier le mouvement dans la bonne direction tout d'abord, la direction initiale k de la rotation est donnée comme $k = 1$ (vers la gauche) si $s(t_0) > 0$ et $k = -1$ (vers la droite) si $s(t_0) < 0$. La commande motrice est alors initialisée à une vitesse angulaire constante et se termine lorsque qu'un changement dans le signe de $s(t)$ est détecté, c'est-à-dire lorsque la valeur d'ILD croise une valeur nulle et que la tête s'est trouvée aligner sur la source.

Notons qu'il n'est pas nécessaire d'introduire des canaux fréquentiels multiples, $s(t)$ pouvant se calculer directement à partir de l'énergie des signaux temporels. Leur présence se justifie par l'utilisation conjointe du comportement d'orientation avec un processus d'échantillonnage de l'espace sensorimoteur au chapitre 7, ce processus nécessitant des canaux multiples. Par ailleurs, une implémentation de ce réflexe basée sur l'ITD est également possible, non plus par minimisation de l'ILD mais par maximisation de l'activité d'ITD autour du délai nul.

6.1.1.2 Localisation *a posteriori*

Considérant l'exécution d'un comportement d'orientation, nous appelons t_0 l'instant initial et t_f l'instant final, obtenus respectivement avant et après minimisation de l'ILD. Soient $s_0 = s(t_0)$, $s_f = s(t_f)$ les états sensoriels initial et final. Soient de plus $m_0 = m(t_0)$ et $m_f = m(t_f)$ les états moteurs associés. Conformément à l'Eq. 5.3 l'état moteur $\tilde{m}$ s'exprime alors *a posteriori* comme l'angle total de rotation effectué durant le mouvement réflexe. Nous avons donc :

$$\tilde{m} = m_f - m_0. \tag{6.2}$$

Puisque nous avons défini l'état sensoriel de référence $\Phi(m_{ref}, e_{ref})$ de l'Eq. 5.3 comme correspondant à la sensation d'une source centrée, sensation qui se retrouve à la fin du réflexe, nous avons de plus :

$$s_f = \Phi(m_{ref}, e_{ref}) + \epsilon_s, \tag{6.3}$$

Dans le cadre d'une expérimentation en milieu anéchoïque et d'une source sonore unique, l'erreur d'estimation ϵ_s peut avoir pour origine l'agent (ego-bruits, précision des moteurs) ou l'environnement (bruits, réverbération). Notons finalement que l'Eq. 6.3 rend possible l'approximation de l'état de référence $\Phi(m_{ref}, e_{ref})$ au travers d'expériences successives du comportement d'orientation. La discussion de ce chapitre revient sur ce point et sur son rôle fondamental dans l'apprentissage autonome de la localisation.

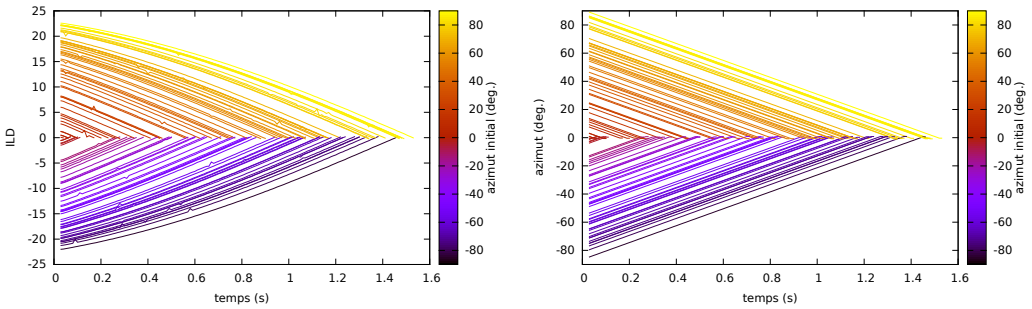
6.1.2 Comportement d'orientation et de déplacement

Ce comportement est une extension directe du comportement d'orientation. En plus de l'orientation de la tête, il permet le déplacement du robot vers l'origine de la source grâce au contrôle d'une paire de roues (Bernard *et al.* , 2010a), d'une manière semblable aux véhicules de Braitenberg (1986).

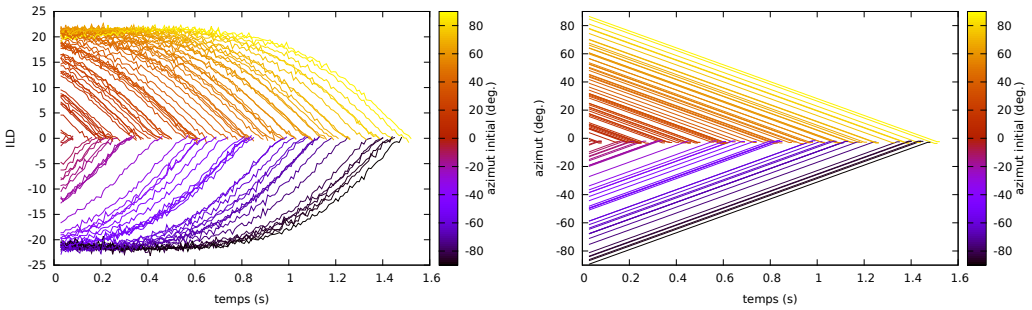
En gardant la même méthodologie que précédemment nous ajoutons un système très simple de contrôle des roues grâce à un vecteur vitesse : la composante de vitesse de ce vecteur est constante tandis que sa composante directionnelle est asservie par l'angle de rotation du cou. Ainsi si le cou est orienté à 45° vers la droite, le robot s'oriente progressivement à 45° vers la droite, jusqu'à ce que le cou reprenne sa position initiale. Le comportement de phonotaxie est considéré comme terminé dès lors que le cou comme les roues ont retrouvé une orientation de 0°.

6.2 Expériences

Ce paragraphe présente des validations expérimentales des deux comportements de localisation active présentés dans ce chapitre. Ces validations sont obtenues en simulation et sur deux plateformes robotiques.



(a) Filtrage h_{dir}. Erreur finale moyenne à $0.54 \pm 0.2°$.



(b) Filtrage h_{hrtf}. Erreur finale moyenne à $2.7 \pm 0.5°$.

FIGURE 6.1 – Détails de la simulation de 100 comportements d'orientation en réponse à des sources d'azimut aléatoire compris entre -90° et 90° (dégradé de couleurs) pour deux modèles d'HRTF présentés au paragraphe 4.1.1 : (a) h_{dir} (modèle 4) et (b) h_{hrtf} (modèle 3). A gauche : évolution de $s(t)$ durant le réflexe. A droite : évolution de l'erreur de localisation $\Delta\tilde{m}$.

6.2.1 Comportement d'orientation

Nous présentons ici successivement une validation du comportement d'orientation en simulation puis sur une plateforme robotique. Nous considérons une variété environnementale $\mathcal{E}$ modélisant un milieu anéchoïque composé d'une source sonore unique stationnaire diffusant un bruit blanc échantillonné à 44.1 kHz. La source est positionnée par rapport à l'agent à une distance fixe et dans l'intervalle azimutal $[-90°, 90°]$ de manière aléatoire et uniforme. La variété motrice $\mathcal{M}$ est constituée du degré de liberté de rotation en azimut offert par le cou. Après l'initiation du réflexe, la vitesse de rotation du cou est fixée à 60 deg.s^{-1}.

Après convergence du réflexe, nous définissons l'erreur de localisation $\Delta\tilde{m}$ comme la valeur absolue de la différence entre l'azimut réel de la source e_ϕ et l'angle de rotation total effectué durant le mouvement $\tilde{m}$. Bien sur cette erreur est à considérer du point de vue extérieur, n'étant en aucun cas accessible au robot. Nous avons ainsi :

$$\Delta\tilde{m} = |e_\phi - \tilde{m}|. \tag{6.4}$$

6.2.1.1 En simulation

En simulation tout d'abord, la Fig. 6.1 présente le résultat d'une simulation de 100 comportements d'orientation. Une paire de bancs de 30 filtres gammatone disposés entre 100 Hz et 8 kHz sont utilisés, avec un seuil $\tau = 0$. Du fait de la stationnarité

de la source et des conditions anéchoïques de la simulation, une durée d'intégration $T_{ild} = 10^{-2}$ s relativement courte est utilisée. La fréquence de la simulation, c'est-à-dire le nombre d'estimation de l'ILD et de mise à jour de l'état moteur par seconde, est fixé à 100 Hz, une fréquence plus faible que la fréquence de sous-échantillonnage proposée par le modèle auditif (égale à $2/T_{ild} = 200$ Hz). Dans le but d'observer l'influence des indices spectraux sur les performances du réflexe, nous utilisons les deux modèles d'HRTF h_{hrtf} et h_{dir} introduits respectivement aux paragraphes 4.1.1.3 et 4.1.1.4. Le filtre h_{hrtf} correspond à des HRTF enregistrées sur un mannequin binaural par pas de 5°, les directions intermédiaires étant traitées par interpolation, et h_{dir} consiste en un filtrage purement directionnel, sans indices spectraux. L'erreur finale est largement indépendante de la position initiale de la source. Le réflexe basé sur h_{dir} atteint une erreur finale $\Delta m = 0.54 \pm 0.2°$ en moyenne tandis qu'elle est de $\Delta m = 2.7 \pm 0.5°$ lorsque le réflexe est associé au filtrage h_{hrtf} . Cependant l'évolution de l'ILD au cours du réflexe est plus linéaire dans le cas h_{dir} que de h_{hrtf}, la rapide diminution de l'ILD en fin de réflexe dans ce second cas étant expliquée par la directivité très marquée de ce filtrage (voir Fig. 4.3). Des résultats comparables, non détaillés ici, ont été obtenus avec des sources provenant des 360° azimutaux, la résolution de l'ambiguïté avant/arrière se faisant alors naturellement de manière active.

6.2.1.2 Sur plateforme robotique

Une validation du réflexe d'orientation sur plateforme robotique à été proposée par Garcia (2013) lors de son stage effectué au sein de notre équipe. Nous résumons ici les résultats de ce travail. La plateforme robotique utilisée est le Binnobot conçu par la société BVS (Fig. 6.2) équipé de deux systèmes pavillons/microphones identiques à ceux utilisés sur la plateforme Psikharpax. L'expérience s'est déroulée dans une salle sourde quasi-anéchoïque. À la différence de l'évaluation effectuée en simulation, l'ILD est calculée ici directement sur l'énergie du signal temporel, sans passer par un filtrage cochléaire. Sur un total de 50 réflexes exécutés, les performances atteignent une précision moyenne de 3°, une précision tout à fait convenable et comparable avec les performances moyennes observées chez l'humain (voir Fig. 2.2). La Fig. 6.3 décrit l'évolution des signaux temporels, des énergies gauche et droite, de l'ILD et de l'angle azimutal du cou durant une expérience de réflexe, à partir d'une source située à 63° à droite. L'ensemble des détails concernant l'implémentation, le protocole expérimental et les résultats sont fournis par Garcia (2013).

6.2.2 Comportement d'orientation et de déplacement

À l'inverse des autres expérimentations présentées dans cette thèse, cette expérience de phonotaxie se base sur le modèle cochléaire de Lyon plutôt que sur des filtres gammatone (Bernard *et al.* , 2010a). Nous utilisons un jeu de 32 canaux fréquentiels. Dans le but d'observer l'influence de la directivité de l'oreille externe sur la phonotaxie, La Fig. 6.4 présente des trajectoires obtenues à partir du robot-rat Psikharpax dans 2 configurations différentes, avec et sans les pavillons artificiels, pour 4 positions de source (en face, à droite, à gauche, derrière). La source est ici un bruit blanc stationnaire.

Dans le cas où le robot est équipé de ses pavillons (Fig. 6.4(a)), nous observons que la trajectoire obtenue est quasiment rectiligne pour une source située en face, à gauche et à droite. Concernant la configuration dans laquelle la source est placée

FIGURE 6.2 – La plateforme robotique Binnobot développée par la société BVS et équipée du système auditif périphérique de Psikharpax. Le robot est également équipé de deux caméras mobiles (non utilisées ici). En arrière-plan, un aperçu de la chambre sourde utilisée pour l'expérience. D'après Garcia (2013).

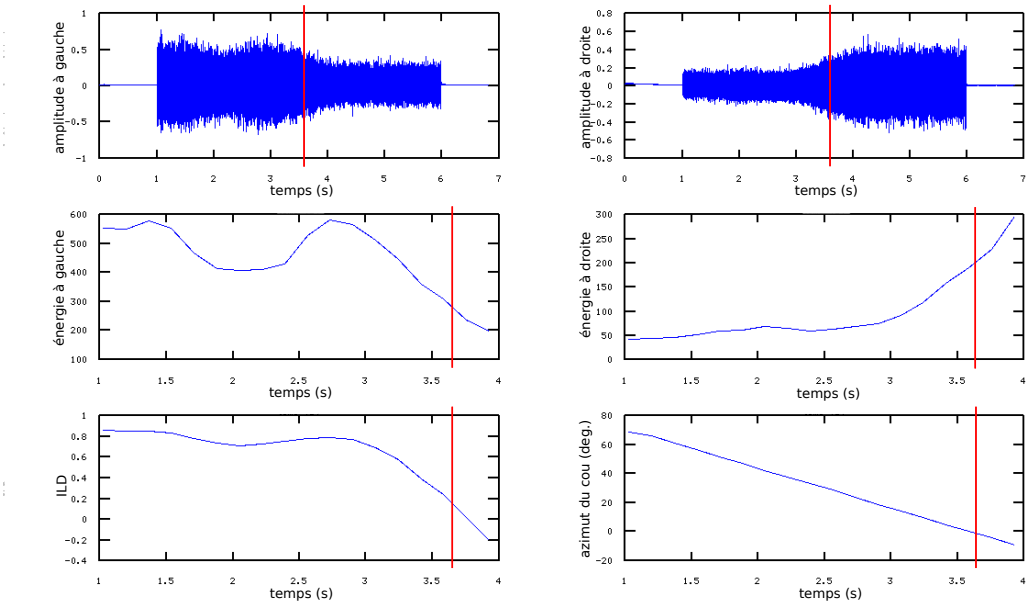
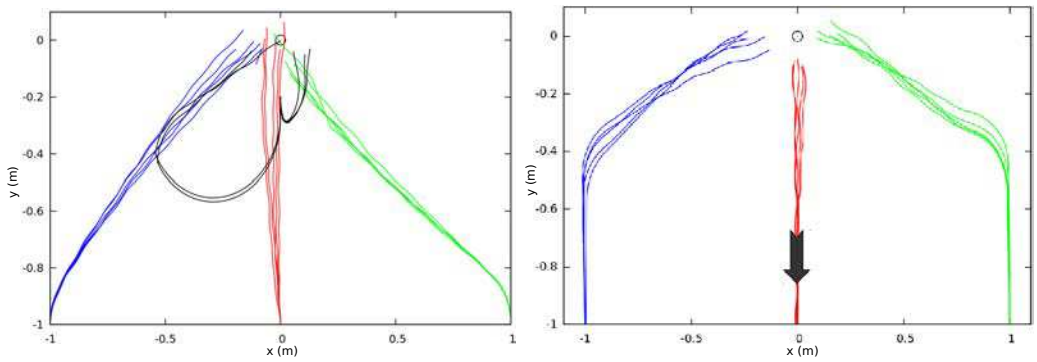


FIGURE 6.3 – Évolution du réflexe d'orientation sur Binnobot à partir d'une source située à 63° à droite. De gauche à droite et de haut en bas : signal temporel gauche, signal temporel droit, énergie à gauche, énergie à droite, ILD, azimut du cou. Les barres verticales rouges représentent l'instant t_f de convergence du réflexe, vers 3.6 s, lorsque l'ILD est annulé et le cou orienté autour de 0°. Adapté de Garcia (2013).



(a) Pavillons installés (b) Pavillons enlevés

Figure 6.4 – Comportement de phonotaxie sur Psikharpax. Trajectoires depuis différentes positions initiales avec les pavillons installés (a) et les pavillons enlevés (b) : source en face (rouge), à gauche (bleu), à droite (vert) et dans le dos (noir). La source sonore, diffusant un bruit blanc et représentée par le cercle noir, est située en $(0, 0)$. 5 trajectoires sont affichées pour chaque configuration. Pour plus de lisibilité, les trajectoires noires de la figure de gauche, toutes superposées, sont représentées par une flèche.

dans le dos du robot, nous observons une trajectoire en demi-tour : la phonotaxie, exploitant implicitement les caractéristiques directionnelles de l'oreille externe, permet de résoudre de manière totalement naïve l'ambiguïté avant-arrière constaté en psychoacoustique (voir à ce sujet le paragraphe 2.1.3). Ainsi une légère perturbation des conditions initiales (le robot n'est pas parfaitement aligné sur la source, les oreilles externes ne sont pas exactement symétriques) provoquent un ILD non nul au départ du robot, expliquant un demi-tour tantôt vers la droite, tantôt vers la gauche. En considérant des conditions initiales parfaitement symétriques et donc un ILD initial nul, il est toutefois possible de conserver un réflexe fonctionnel, par exemple en faisant suivre la détection d'un ILD nul d'une légère rotation du cou permettant d'amorcer le réflexe dans une ou l'autre direction selon le cas.

La Fig. 6.4(b) illustre les résultats obtenus dans les mêmes conditions que précédemment mais avec les pavillons enlevés et les microphones orientés cette fois-ci vers l'avant du robot. La perte de directivité que cette modification entraîne à pour conséquence une moindre performance du comportement de phonotaxie. Ainsi pour les configurations dans lesquelles la source est située sur la droite ou sur la gauche, le robot commence par se diriger tout droit, signifiant qu'il ne perçoit pas un ILD suffisant pour tourner le cou. Cet ILD n'est en effet perçu qu'en fin de trajectoire, lorsque le robot est suffisamment proche de la source. Notons enfin que sans pavillons, la résolution de l'ambiguïté avant-arrière observée précédemment n'est plus constatée, en raison là encore d'un manque de directivité des oreilles externes.

6.3 Discussion

Cette discussion revient tout d'abord sur les capacités de localisation des deux comportements réflexes étudiés. Enfin nous verrons que le comportement d'orientation constitue un point de départ à l'apprentissage autonome d'une capacité de localisation passive qui sera l'objet du prochain chapitre.

Localisation active et naïve Le comportement d'orientation fournit donc une capacité de localisation active à un agent totalement naïf sur son environnement et son propre système moteur. Les capacités d'action se substituent en effet aux connaissances *a priori*. La localisation s'effectue alors *a posteriori*, une fois le réflexe convergé.

Ce modèle, nous l'avons vu, offre de bons résultats, à la fois en simulation et sur plateforme robotique. Néanmoins les conditions expérimentales utilisées sont très favorables à la tâche de localisation : une source unique et stationnaire, diffusée dans un milieu anéchoïque. Nous considérons en effet le traitement des conditions adverses comme un problème posé non pas directement au comportement d'orientation mais au modèle auditif sur lequel il est basé. Le chapitre 4 a ainsi montré que l'ILD est robuste aux conditions échoïques et à la non-stationarité de la source, sous réserve d'une paramétrisation du modèle auditif adaptée au contexte environnemental. Ainsi les bonnes conditions expérimentales constituent ici une simplification pratique, et non une limitation théorique de la méthode proposée.

Le paragraphe 6.2.2 a de plus démontré que l'utilisation du comportement d'orientation en association avec un filtrage périphérique adapté permet de résoudre l'ambiguïté avant/arrière de manière implicite et totalement naïve. Notons finalement qu'il est possible d'inverser la « polarité » du comportement réflexe, à la manière des véhicules de Braitenberg (1986) (voir Fig. 3.3), le réflexe devenant alors un comportement d'éloignement. Il suffit pour cela non plus de minimiser mais de maximiser l'ILD en inversant simplement le sens de rotation du cou.

Vers l'apprentissage de la localisation Le comportement d'orientation est présent chez l'humain. Il se retrouve en effet dès les premiers stades de développement du nourrisson (Kearsley, 1973), suggérant ainsi qu'il ait un caractère inné et « câblé en dur ». De plus ce comportement semble impliqué dans les phases ultérieures d'apprentissage de la localisation de sources sonores (Muir *et al.* , 1989; Metta, 2000).

L'Eq. 6.3 nous montre ainsi que le réflexe d'orientation, en plus de fournir une estimation *a posteriori* de l'état moteur $\tilde{m}$, nous permet d'accéder de manière totalement naïve à une approximation de l'état sensoriel de référence $\Phi(m_{ref}, e_{ref})$. Selon notre définition de la localisation, la connaissance de cette sensation de référence est fondamentale pour l'estimation de $\tilde{m}$ (Eq. 5.3). Son apprentissage par l'intermédiaire du comportement d'orientation permet donc de réduire la dépendance du modèle à l'environnement.

Ce réflexe est ainsi utilisé par le processus d'apprentissage autonome de la localisation présenté au chapitre 7, à la fois pour l'apprentissage de l'état sensoriel de référence et pour l'auto-supervision de la localisation.

Chapitre 7

Apprentissage autonome de la localisation

Sommaire

Le chapitre 5 a proposé une méthode de localisation sans déplacement qui nécessite la préexistence d'un échantillonnage de l'espace sensorimoteur. A l'inverse le chapitre 6 a proposé un modèle de perception active offrant une forme primitive de localisation, indissociable du déplacement de l'agent vers la source sonore, mais ne nécessitant pas d'échantillonnage ou toute autre connaissance *a priori*. Ce chapitre propose une méthode d'apprentissage de la localisation azimutale qui fusionne ces deux modèles. Plus précisément le comportement d'orientation est utilisé pour apprendre de manière autonome et en ligne, c'est-à-dire expérience après expérience, un échantillonnage de l'espace sensorimoteur qui est ensuite utilisé pour une localisation passive. Après une localisation passive, un processus d'auto-supervision permet également à l'agent de vérifier activement la qualité de l'estimation et, si nécessaire, de la corriger par un comportement d'orientation.

Nous avons vu de plus au chapitre 4 que la majorité des modèles auditifs robotiques intègrent l'action comme l'étape terminale d'un processus purement passif. L'approche sensorimotrice suivie ici place au contraire la perception active comme une condition préalable à la perception sans déplacement. L'expérience sensorimotrice nécessaire à la localisation passive, qui nous imposait une approche supervisée au chapitre 5, est ici directement fournie par les facultés de perception active de

bas niveau. Ceci nous permet donc de proposer une méthode de localisation passive indépendante de toute connaissance de l'agent sur son environnement et son propre système moteur.

Ce chapitre détaille tout d'abord la méthode d'apprentissage avant de présenter une série de résultats expérimentaux concernant l'évolution de l'apprentissage et les performances obtenues en termes de localisation. Cette méthode et les résultats proposés sont en partie tirés de Bernard *et al.* (2012).

7.1 Formalisation

Ce paragraphe présente en détail l'algorithme d'apprentissage autonome de la localisation. La construction de l'échantillonnage sensorimoteur est tout d'abord introduite, suivie par le processus de localisation sur cet échantillonnage et enfin le processus d'auto-supervision.

Notons enfin que la méthode présentée ici repose sur le comportement d'orientation proposé au chapitre 6, ce qui limite son application à la localisation azimutale. Nous reviendrons sur ce point en discussion.

7.1.1 Construction active de l'échantillonnage

Soient un espace environnemental $\mathcal{E}$, un espace moteur $\mathcal{M}$ et un espace sensoriel $\mathcal{S}$. Un échantillonnage de taille n de l'espace sensorimoteur est défini comme l'ensemble (S_0, S_f, M, A), où $S_0 \in \mathcal{S}^n$ et $S_f \in \mathcal{S}^n$ correspondent respectivement aux ensembles des états sensoriels initiaux et finaux obtenus avant et après le comportement d'orientation, où $M \in \mathcal{M}^n$ est l'ensemble des déplacements moteurs effectués durant les expériences de réflexe et où $A \in (\mathcal{S} \times \mathcal{M})^n$ est la carte sensorimotrice associant chaque état sensoriel initial au changement d'état moteur opéré durant un réflexe (voir le paragraphe 6.1.1.2).

Au chapitre 5 cet échantillonnage était donné *a priori*. Nous nous plaçons maintenant dans un contexte non-supervisé où l'échantillonnage est appris de manière itérative. Chaque exécution d'un comportement d'orientation donne en effet accès à un nouvel échantillon. Supposons ainsi que l'agent expérimente un état sensoriel $s_0 = \Phi(m_0, e)$ lié à l'état moteur $m_0 \in \mathcal{M}$ et à un état environnemental constant $e \in \mathcal{E}$. L'exécution du comportement d'orientation, qui mène à l'état moteur m_f donne accès à l'état sensoriel final $s_f = \Phi(m_f, e)$ et au déplacement moteur $m_f - m_0$ effectué. Après l'exécution du comportement, l'échantillonnage (S_0, S_f, M, A) est mis à jour, de sorte que :

$$\begin{array}{rcl} S_0 & \leftarrow & S_0 \cup \{s_0\}, \\ S_f & \leftarrow & S_f \cup \{s_f\}, \\ M & \leftarrow & M \cup \{m_f - m_0\}, \\ A & \leftarrow & A \cup \{(s_0, m_f - m_0)\}. \end{array} \tag{7.1}$$

Après n expériences de comportement d'orientation nous avons donc $S_0 = \{s_{0,i}\}_{i\in[1,n]}$, $S_f = \{s_{f,i}\}_{i\in[1,n]}$, $M = \{m_{f,i} - m_{0,i}\}$ et $A = \{(s_{0,i}, m_{f,i} - m_{0,i})\}_{i\in[1,n]}$.

L'agent étant initialement totalement naïf à son espace sensorimoteur, nous avons initialement $S_0 = S_f = M = A = \emptyset$. La taille minimale de S_0 doit cependant être égal à l'ordre du voisinage k pour effectuer une interpolation. Une phase d'initialisation est donc ajoutée, de sorte que les $2k$ premières expériences sont systématiquement traitées par le comportement d'orientation.

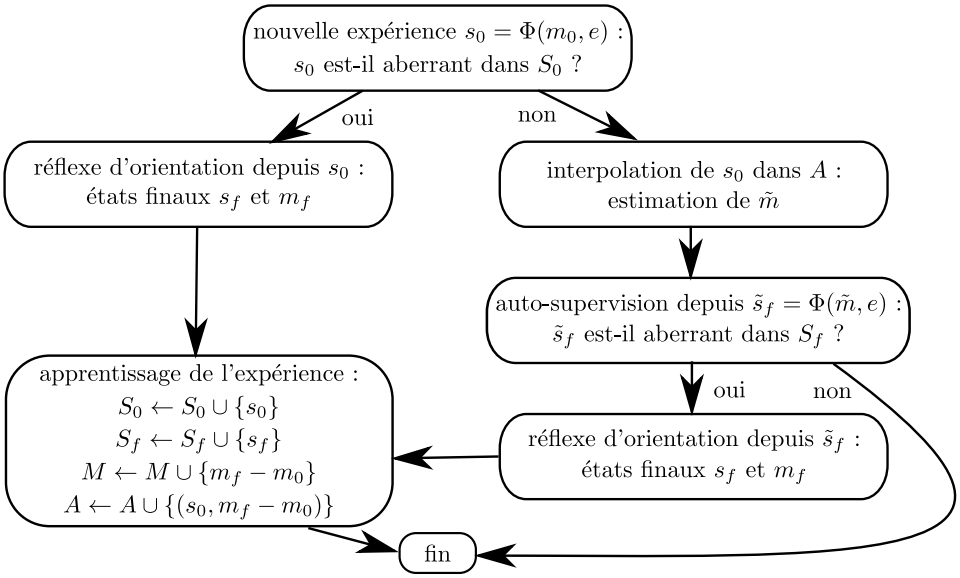


FIGURE 7.1 – Automate représentant l'algorithme d'apprentissage autonome de la localisation azimutale. Voir le paragraphe 7.1 pour le détail de chacune des étapes.

7.1.2 Localisation passive

Supposons que l'échantillonnage sensorimoteur (S_0, S_f, M, A) soit de taille $n > k$, de sorte qu'une localisation passive soit possible par interpolation. Soit également $s_0 = \Phi(m_0, e)$ un état sensoriel expérimenté par l'agent. Pour que l'interpolation s'effectue correctement, le voisinage de s_0 dans S_0 doit être suffisamment dense. Considérant cependant l'ensemble S_0 composé de peu de points, ou bien un état sensoriel s_0 très différent de ceux précédemment expérimentés et inclus dans S_0, s_0 peut se situer dans une zone de S_0 sans aucun voisin proche. Cette situation mène donc à une interpolation de $\tilde{m}$ erronée ou, pour le moins, peu fiable.

Pour repérer ces états sensoriels isolés, s_0 est confronté à S_0 par un test statistique détectant les données aberrantes, et deux issues peuvent alors se présenter. Si s_0 est détecté comme aberrant dans S_0, l'interpolation n'est pas effectuée. À la place, le comportement d'orientation est lancé pour une localisation active et l'échantillonnage sensoriel est alors mis à jour avec cette expérience supplémentaire (Eq. 7.1). Si au contraire s_0 n'est pas aberrant, c'est-à-dire s'il a des voisins proches dans S_0, sa position est estimée par interpolation dans ce voisinage (Eq. 5.4). Après cette estimation passive, un processus d'auto-supervision permet de vérifier la qualité de l'interpolation et de la corriger de manière active si nécessaire.

L'algorithme complet intégrant ces différents éléments est représenté sous la forme d'un automate par la Fig. 7.1. Ce processus est exécuté pour chaque nouvel état sensoriel s_0 expérimenté par l'agent. Ce paragraphe présente maintenant le détail de la méthode de détection des données aberrantes et du processus d'auto-supervision.

7.1.2.1 Détection des états sensoriels aberrants

La détection des états sensoriels aberrants repose sur la notion de distance sur l'espace sensoriel et nécessite l'introduction d'une métrique $d : \mathcal{S} \times \mathcal{S}^n \to \mathbb{R}^+$. Puisque cette métrique est utilisée dans un contexte d'interpolation aux k-voisins, nous utilisons une métrique d_{kppv} correspondant à la distance moyenne entre un état s_0 et ses

k voisins dans l'ensemble S_0. Nous avons ainsi :

$$d_{kppv}(s_0, S_0) = \frac{1}{k} \sum_{s_i \in K_{S_0}} ||s_0 - s_i||, \tag{7.2}$$

où K_{S_0} est l'ensemble des k-ppv de s_0 dans S_0, comme défini au paragraphe 5.2.3.2.

À partir de cette métrique, la détection des états sensoriels aberrants est effectuée par le test de Grubbs (1969), également appelé le test résiduel normé maximum. Ce test statistique, qui assume des données provenant d'une distribution normale, compare la valeur de $d_{kppv}(s_0, S_0)$ avec la valeur moyenne de cette métrique sur tous les points de S_0. Plus précisément, s_0 est considéré comme aberrant dans S_0 si :

$$\frac{|d(s_0, S_0) - \mu(S_0)|}{\sigma(S_0)} > v_{crit}, \tag{7.3}$$

où $\mu(S_0)$ et $\sigma(S_0)$ correspondent respectivement à la moyenne et à l'écart-type des valeurs de la métrique d_{kppv} pour tous les points $s_i \in S_0$. Enfin v_{crit} est une valeur critique qui dépend du nombre n de points dans S_0 et de deux paramètres α et β. Elle s'exprime en fonction de n, α et β comme :

$$v_{crit} = \frac{\beta(n-1)}{\sqrt{n}} \sqrt{\frac{(t^{n-2}_{\alpha/2n})^2}{n-2+(t^{n-2}_{\alpha/2n})^2}}, \tag{7.4}$$

où $t^{n-2}_{\alpha/2n}$ représente l'inverse de la fonction de répartition d'une loi de Student à $n-2$ degrés de liberté pour une probabilité $\alpha/2n$ (Shaw, 2006). Cette valeur α permet de fixer la sensibilité de détection des données aberrantes (Grubbs, 1969). Néanmoins le paramètre α modifie la sensibilité du test de façon trop marginale. Nous avons donc introduit le paramètre β, qui permet une mise à l'échelle de la valeur critique et un réglage plus efficace de la sensibilité du test de Grubbs sur les espaces sensoriels étudiés (voir le paragraphe 7.2.2).

7.1.2.2 Auto-supervision

Supposons un état sensoriel $s_0 = \Phi(m_0, e)$ considéré comme non-aberrant dans S_0 (Eq. 7.3). Comme expliqué ci-dessus, l'état moteur $\tilde{m}$ estimant la position de la source sonore est alors obtenu par interpolation (Eq. 5.4). Le processus d'auto-supervision permet alors de vérifier la qualité de l'estimation $\tilde{m}$ et de la corriger si nécessaire. Pour cela l'agent s'oriente vers la direction estimée et compare l'état sensoriel $\tilde{s}_f = \Phi(\tilde{m}, e)$ obtenu avec l'état sensoriel de référence $\Phi(m_{ref}, e_{ref})$ nécessaire à la localisation sans déplacement (voir le paragraphe 6.1.1.2). Puisque l'ensemble S_f regroupe les états sensoriels obtenus après déplacement, sa moyenne $\mu(S_f) = \frac{1}{|S_f|} \sum_{s_i \in S_f} s_i$ permet d'estimer cet état de référence.

Après un mouvement de l'agent en $\tilde{m}$, l'état sensoriel $\tilde{s}_f = \Phi(\tilde{m}, e)$ est alors comparé à l'ensemble S_f par un test de Grubbs. Puisque l'état sensoriel de référence est estimé par la moyenne $\mu(S_f)$, nous utilisons ici la métrique $d_{moy}(\tilde{s}_f, S_f)$ telle que :

$$d_{moy}(\tilde{s}_f, S_f) = ||\tilde{s}_f - \mu(S_f)||, \tag{7.5}$$

où $\mu(S_f)$ correspond à la moyenne des éléments $s_f \in S_f$. Si $\tilde{s}_f$ est détecté comme étant aberrant dans S_f par la métrique, cela signifie que cet état sensoriel est trop éloigné de l'état de référence et donc que l'estimation $\tilde{m}$ était erronée. Dans ce cas la correction de l'erreur s'effectue par l'exécution du comportement réflexe et la mise à jour de la carte sensorimotrice est opérée comme précédemment (Eq. 7.1).

7.2 Expérience

Ce paragraphe présente une série de résultats détaillant l'apprentissage autonome de la tâche de localisation azimutale. Le protocole expérimental est détaillé, puis les résultats d'apprentissage sont présentés. Il est tout d'abord montré que l'état sensoriel de référence peut être approximé par des expériences successives de comportement d'orientation. L'évolution de l'apprentissage est ensuite analysée, il est montré que 200 itérations environ sont nécessaires pour que la localisation passive atteigne ses performances optimales. Enfin l'influence de la sensibilité des tests de Grubbs pour la détection des données aberrantes est évaluée.

7.2.1 Protocole expérimental

L'environnement est modélisé par un espace bidimensionnel anéchoïque dans lequel une source unique et omnidirectionnelle émet un bruit blanc échantillonné à 44.1 kHz. L'azimut de la source est choisi aléatoirement dans l'intervalle $[-90, 90]$ selon une distribution uniforme. La variété environnementale s'exprime donc comme $\mathcal{E} = [-90, 90]$.

L'agent est situé à la position fixe $(0, 0)$ et possède un degré de liberté lui permettant une rotation de la tête dans le plan azimutal. La vitesse de rotation du cou est fixée à 60 deg.s^{-1}. Nous notons $m_\phi \in [-90, 90]$ l'orientation de la tête, qui est modélisée comme un axe binaural de distance interaurale fixée à 0.19 m auquel est ajouté le modèle d'oreilles externes (h^L_{dir}, h^R_{dir}) constitué d'un filtrage purement directionnel (paragraphe 4.1.1.4). La distance interaurale est donnée ici à titre purement indicatif, sa connaissance n'étant pas nécessaire. De plus le seul degré de liberté accessible à l'agent est m_ϕ, ce qui nous donne ainsi directement l'expression de la variété motrice sous-jacente $\mathcal{M} = [-90, 90]$.

Le système auditif est constitué d'une paire de $n = 30$ filtres gammatone disposés de 100 Hz à 8 kHz. Le seuil de transduction τ est fixé à 0. Seul l'ILD est considéré et la durée d'intégration T_{ild} est fixée à 20 ms. La variété sensorielle $\mathcal{S}$ est ici constituée des vecteurs d'ILD générés par le modèle auditif, nous avons donc $\mathcal{S} = [-1, 1]^{30}$. Enfin, les paramètres de l'algorithme sont fixés à leur valeur par défaut et nous avons $k = 12$, $\alpha_{kppv} = 0.05$, $\beta_{kppv} = 1$, $\alpha_{moy} = 0.05$ et $\beta_{moy} = 1$.

7.2.2 Résultats

Ce paragraphe présente tout d'abord une analyse de l'échantillonnage sensorimoteur obtenu après apprentissage, puis revient sur le déroulement de cet apprentissage. Enfin l'influence du processus de détection de données aberrantes sur l'apprentissage et la localisation est détaillé.

7.2.2.1 Analyse de l'échantillonnage sensorimoteur

Avant d'analyser le détail du déroulement de l'algorithme, intéressons-nous ici à l'échantillonnage obtenu après 100 comportements d'orientation exécutés à partir d'autant de configurations environnementales différentes. La Fig. 7.2 propose une représentation 2D de cet échantillonnage. La représentation de S_0 est à rapprocher de la variété obtenue à partir du comportement d'orientation présentée Fig. 5.5(a). On retrouve en effet dans les deux cas une topologie parabolique, à l'exception que la tâche de localisation n'est pas ici confrontée à l'ambiguïté avant/arrière. La projection des états sensoriels finaux S_f sur la représentation de S_0 est également présentée

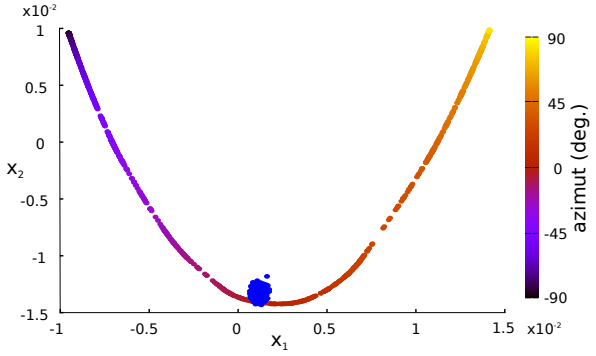


FIGURE 7.2 – Représentation 2D des échantillonnages S_0, S_f et M obtenus après 100 comportements d'orientation. Les points reposant sur la parabole sont les états sensoriels initiaux dans S_0 et sont estimés par l'algorithme des cartes propres Laplaciennes (voir l'annexe C.1.2). Le dégradé de couleurs représente M et correspond à l'azimut de la source. Enfin les points bleus représentent les états sensoriels finaux dans S_f après projection sur la parabole (la méthode de projection est présentée dans l'annexe C.1.3).

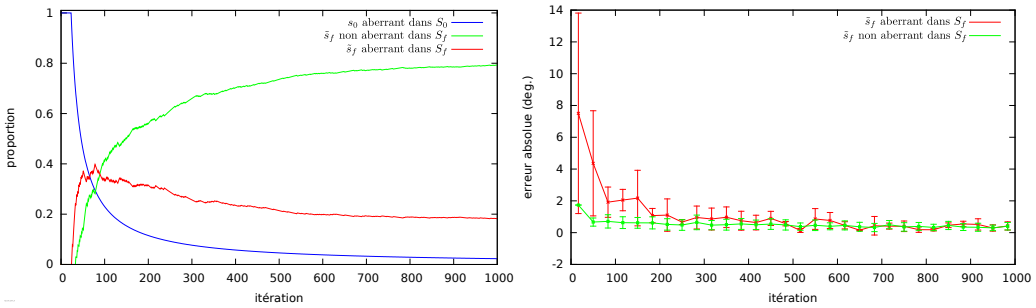
sur la Fig. 7.2. Ces états finaux sont tous projetés dans une même région de la variété indiquant une source sonore centrée.

Ce résultat consiste donc en une validation expérimentale de l'existence de l'état sensoriel de référence $\Phi(m_{ref}, s_{ref})$ tel que proposé par l'Eq. 5.3. Il est de plus montré que le comportement d'orientation permet d'estimer cet état de référence. Enfin il est constaté que S_f possède une topologie en disque, confirmant le choix de la métrique d_{moy} retenue pour la détection des états aberrants, de même que la topologie de S_0 motive le choix de la métrique d_{kppv} (voir le paragraphe 7.1.2).

7.2.2.2 Évolution de l'apprentissage

La Fig. 7.3 détaille les résultats obtenus durant l'exécution de l'algorithme sur 1000 itérations, chacune étant associée à une source sonore d'azimut aléatoire. Après 1000 itérations, 2.4% des expériences ont été considérées comme aberrantes dans S_0 et se sont conclues par un comportement d'orientation. Celles-ci correspondent aux $2k = 24$ premières expériences. Les autres expériences sont localisées passivement, le processus d'auto-supervision étant alors engagé à partir de l'état sensoriel $\tilde{s}_f = \Phi(\tilde{m}, e)$. Parmi les 1000 expériences, 79% ont vu l'estimation $\tilde{m}$ acceptée par l'auto-supervision, c'est-à-dire que l'état $\tilde{s}_f$ associé fut non-aberrant dans S_f. Enfin 18% ont vu leur interpolation rejetée, c'est-à-dire avec $\tilde{s}_f$ aberrant dans S_f. Il est également montré que le processus d'auto-supervision permet de détecter les estimations associées aux erreurs de localisation les plus importantes (1.5° contre 0.4° en moyenne). Il est de plus constaté que la méthode d'interpolation permet d'atteindre une erreur moyenne inférieure à l'erreur moyenne obtenue par le comportement d'orientation (0.5° lorsque $\tilde{s}_f$ est non-aberrant dans S_f, contre 0.9° pour le réflexe).

La Fig. 7.3(a) présente l'évolution de la répartition des 3 issues possibles du processus de localisation : exécution du comportement d'orientation (*i.e.* s_0 est aberrant dans S_0), auto-supervision validée ou corrigée ($\tilde{s}_f$ est respectivement non-aberrant ou aberrant dans S_f). La Fig. 7.3(b) détaille l'erreur de localisation associée aux cas ayant entraîné une localisation passive, selon que l'estimation ait été validée ou corri-



(a) Évolution de l'apprentissage (b) Erreur absolue de localisation

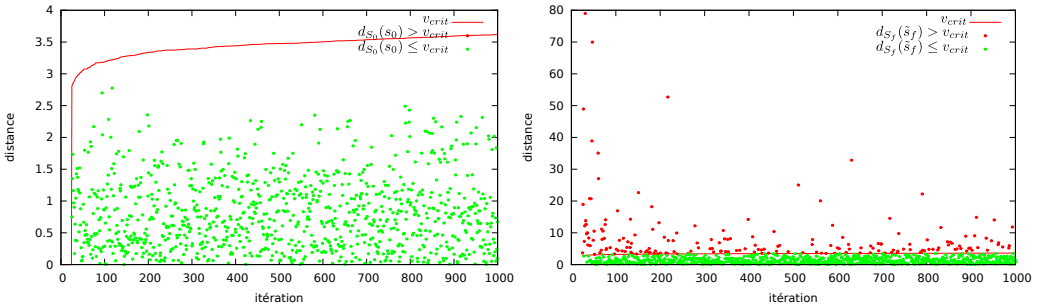
Figure 7.3 – Déroulement de l'apprentissage sur 1000 itérations. (a) Évolution de la proportion des expériences traitées par le comportement d'orientation (en bleu, erreur moyenne finale à $0.9 \pm 0.1°$), des expériences localisées par interpolation dont l'estimation est acceptée par l'auto-supervision (en vert, erreur moyenne finale à $0.4 \pm 0.3°$) ou au contraire détectée comme erronée puis corrigée par le comportement d'orientation (en rouge, erreur moyenne finale à $1.5 \pm 2.4°$). (b) Erreur absolue obtenue au cours de l'apprentissage selon l'issue de l'auto-supervision (moyenne et écart-type pour 33 itérations successives).

gée par l'auto-supervision. À partir de ces données nous pouvons distinguer 3 phases dans l'apprentissage. Une phase d'initialisation premièrement, qui concerne les 24 premières itérations et pour lesquelles le comportement d'orientation est systématiquement exécuté. Cette phase correspond à la phase d'initialisation mise en place pour garantir une taille minimale à l'échantillonnage avant la première interpolation (voir le paragraphe 7.1.1). La seconde phase consiste en une phase d'apprentissage de l'espace auditif et intervient jusqu'à l'itération 200 environ. Durant cette période, le nombre d'estimations de $\tilde{m}$ rejetées par l'auto-supervision est élevé, de même que l'erreur de localisation associée. Enfin la dernière phase intervient lorsque l'apprentissage a convergé, c'est-à-dire lorsque l'erreur de localisation se stabilise et que 80% environ des expériences sont traitées passivement et acceptées par l'auto-supervision. Ce nombre d'environ 200 itérations nécessaires à l'apprentissage d'un demi espace azimutal est cohérent avec les 500 échantillons nécessaires pour obtenir de bonnes performances lors de l'apprentissage de l'espace azimutal complet, comme illustré au paragraphe 5.3.1.

7.2.2.3 Détection des états sensoriels aberrants

Les deux détections d'états sensoriels aberrants sont des étapes importantes du processus de localisation, puisqu'elles en déterminent l'issue. Ce paragraphe revient sur le détail de ces détections, particulièrement sur l'évolution des valeurs critiques associées aux 2 tests de Grubbs.

Ainsi la Fig. 7.4(a) détaille la détection des états sensoriels aberrants dans S_0, qui exploite la métrique d_{kppv}. Nous voyons ici qu'aucun état sensoriel n'a été considéré aberrant dans S_0, toutes les distances étant inférieures à la valeur critique. Ce résultat s'explique par les conditions expérimentales retenues (conditions anéchoïques, source unique, spectre stationnaire) qui génèrent des états sensoriels dont la variabilité n'est pas suffisante pour que la distance d_{kppv} associée dépasse la valeur critique. Néanmoins cette détection constitue un garde-fou : la perception d'un ton pur par



(a) 1ère détection de données aberrantes (d_{kppv}) (b) 2ème détection de données aberrantes (d_{moy})

FIGURE 7.4 – Détection des données aberrantes et évolution de la valeur critique sur les 1000 itérations de la Fig. 7.3. (a) Évolution de la valeur critique associée à S_0 et valeurs des distances $d_{kppv}(s_0, S_0)$ centrées réduites (voir Eq. 7.3). (b) Évolution de la valeur critique associée à S_f et valeurs des distances $d_{moy}(\tilde{s}_f, S_f)$ centrées réduites.

exemple engendrerait un état sensoriel très différents de ceux de S_0, dont seul un canal fréquentiel serait non-nul, provoquant ainsi une distance d_{kppv} élevée et une détection de cet état comme aberrant dans S_0.

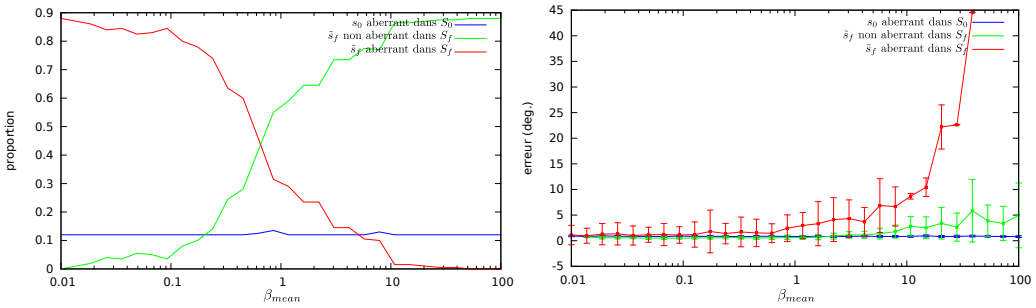
Deuxièmement la Fig. 7.4(b) détaille la détection des données aberrantes dans S_f, associée à l'auto-supervision et à la métrique d_{moy}. Nous remarquons tout d'abord la stabilité de la valeur critique au cours de l'apprentissage. Cette stabilité est obtenue par l'utilisation de d_{moy} qui, basée sur la distance à la moyenne de S_f, est insensible à la densité de l'échantillonnage et à ses caractéristiques locales. Cette stabilité de la valeur critique a pour conséquence un rejet systématique d'environ 20% des états sensoriels finaux (voir Fig. 7.4), même lorsque l'apprentissage a convergé.

7.2.2.4 Sensibilité de l'auto-supervision

Les résultats présentés ci-dessus ont été obtenus à partir de paramètres constants. Les paramètres associés aux tests de Grubbs ont notamment été fixés a $\alpha_{kppv} = \alpha_{moy} = 0.05$ et $\beta_{kppv} = \beta_{moy} = 1$. Nous avons vu que, après convergence de l'apprentissage, l'auto-supervision rejette 20% environ des états sensoriels. Cette expérience se propose d'évaluer l'influence du paramètre β_{moy} sur la détection des données aberrantes dans S_f et donc sur la sensibilité de l'auto-supervision. La Fig. 7.5 présente ainsi les conditions finales obtenue après 1000 itérations selon différentes valeurs de β_{moy} variant de 10^{-2} à 10^2. Ce paramètre est un simple facteur multiplicatif de la valeur critique qui permet de moduler la sensibilité du test de Grubbs aux données aberrantes (voir Eq. 7.3).

Ainsi la Fig. 7.5(a) détaille la proportion d'expériences associées à chacune des 3 issues possibles de l'algorithme selon la valeur de β_{moy}. La proportion de réflexes reste évidemment constante car non concernée par la détection dans S_f. Il est ainsi constaté qu'une faible valeur de β_{moy} entraîne une sélectivité très élevée (tous les $\tilde{s}_f$ sont aberrants dans S_f pour $\beta_{moy} = 10^{-2}$), le phénomène inverse étant observé lorsque β_{moy} est important (aucun $\tilde{s}_f$ n'est aberrant dans S_f pour $\beta_{moy} = 10^2$).

La Fig. 7.5(b) représente l'erreur de localisation moyenne obtenue en fonction de β_{moy}. Les résultats intuitivement attendus y sont observés. Ainsi l'erreur moyenne associée aux états sensoriels aberrants croît avec l'augmentation de β_{moy}, indiquant que seules les plus larges erreurs de localisation sont rejetées à mesure que la sensibilité



(a) Conditions finales de l'algorithme selon la sensibilité de l'auto-supervision

(b) Erreur moyenne finale associée

Figure 7.5 – Conditions finales de l'algorithme après 1000 itérations en fonction de la sensibilité β_{moy} associée à la détection des données aberrantes dans S_f. β_{moy} varie de 0.01 à 100 selon une échelle logarithmique. (a) proportions d'expériences associées à un comportement d'orientation (bleu), associées à une auto-supervision validée (vert) ou corrigée (rouge). (b) erreur moyenne $\Delta\tilde{m}$ obtenue pour les expériences associées à un comportement d'orientation (bleu), associées à une auto-supervision validée (vert) ou corrigée (rouge).

du test décroît. Là encore l'erreur associée aux mouvements réflexes reste constante puisqu'elle est indépendante du processus d'auto-supervision. Finalement le paramètre β_{moy} permet donc de régler la sensibilité du test associé à l'auto-supervision et de fixer un compromis entre 2 extrêmes : un test très sensible donnant des résultats précis mais reposant lourdement sur le réflexe, ou un test plus lâche mais donnant une erreur plus importante.

7.3 Discussion

L'algorithme introduit dans ce chapitre intègre donc le comportement d'orientation et la localisation passive au sein d'un même modèle, permettant ainsi l'apprentissage autonome de la tâche de localisation dans le plan azimutal. Contrairement à la majorité des méthodes proposées dans l'état de l'art, cette approche considère la localisation active comme une condition préalable à l'apprentissage de la localisation sans déplacement. Cette discussion revient sur les aspects autonomes de ce modèle, puis sur ces limitations. Enfin l'opportunité de développer des stratégies d'explorations indépendantes du comportement d'orientation est envisagée.

Apprentissage autonome Cet algorithme d'apprentissage de la localisation passive peut être qualifié d'autonome pour deux raisons. Premièrement de part son fonctionnement itératif et auto-supervisé, et deuxièmement parce qu'il ne repose sur aucune connaissance *a priori* de l'environnement. Ainsi cet algorithme peut fonctionner dans différentes conditions environnementales et apprendre à localiser différents spectres, à la seule condition que le comportement d'orientation donne des résultats satisfaisants dans ces conditions.

Plus précisément les *a priori* requis par l'algorithme sont les suivants : un système auditif périphérique et un réflexe d'orientation adaptés à l'environnement (robuste à des conditions acoustiques potentiellement adverses), une métrique sur l'espace sensoriel permettant le calcul des k-ppv et la détection des états aberrants, et enfin

une « mémoire » permettant de stocker l'échantillonnage de l'espace sensorimoteur. L'utilisation de l'ITD en lieu et place de l'ILD aurait de plus nécessité la connaissance de la distance interaurale de l'agent.

De plus, à l'exception des paramètres propres au système auditif, un total de 5 paramètres sont requis par cette méthode : l'ordre de voisinage k et les paramètres liés aux détections des données aberrantes : α_{kppv} et β_{kppv} associés au premier test de Grubbs, α_{moy} et β_{moy} associés au second. Si les variations de α_{kppv} et α_{moy} n'influe que de manière marginale sur la détection, nous avons vu que le β_{moy} permet de régler la sensibilité de l'auto-supervision.

Limitations Cet algorithme est présenté dans ce chapitre sous une forme simple et plusieurs améliorations devraient accroître significativement ses performances. La carte sensorimotrice est construite à partir des états sensoriels initiaux uniquement, ainsi l'ajout des états intermédiaires et finaux devrait permettre d'augmenter la quantité de donnée acquise à chaque expérience et en conséquence de réduire le nombre d'expériences auditives requises pour la construction de la carte sensorimotrice. Deuxièmement l'estimation de l'état moteur $\tilde{m}$ est effectuée par l'algorithme comme une simple interpolation en fonction de la distance du voisinage tandis que, autour de chaque point, la variété s'étend selon une direction privilégiée. Ainsi l'ajout d'un facteur directionnel dans la fonction d'interpolation devrait améliorer les performances de l'estimation (Shepard, 1968). Notons enfin là encore la simplicité des conditions expérimentales utilisées. Il serait intéressant d'évaluer cet algorithme sur des tâches plus complexes faisant intervenir des spectres ou des conditions acoustiques variées, ce qui permettrait de qualifier son potentiel de généralisation.

Vers l'exploration de l'espace sensorimoteur Prenons pour conclure un peu de distance par rapport à l'approche présentée dans ce chapitre. Premièrement le modèle présenté ici diffère des approches proposées par Philipona *et al.* (2003) et Laflaquière (2013). L'utilisation du réflexe et du système auditif périphérique centre en effet l'apprentissage sensorimoteur sur une tâche précise, celle de la localisation en azimut, et n'a pas vocation à la découverte de propriétés générales de l'espace physique. Si la découverte de telles propriétés est d'un intérêt théorique évident, nous avons montré dans ce chapitre que la théorie sensorimotrice peut être appliquée avec profit dans le cadre de la perception autonome de sources sonores. Deuxièmement, le comportement d'orientation ayant été validé sur plateforme robotique et la méthode d'interpolation utilisée sur des signaux auditifs réalistes, l'algorithme d'apprentissage proposé semble adapté au contexte de la robotique autonome. Néanmoins, pour dépasser la dépendance de l'apprentissage au comportement d'orientation, il serait intéressant d'envisager d'autres stratégies d'exploration, basées sur une découverte aléatoire ou volontaire de l'espace sensorimoteur. Nous envisagerons ainsi une stratégie d'exploration motivée par la minimisation d'une incertitude perceptuelle au chapitre suivant.

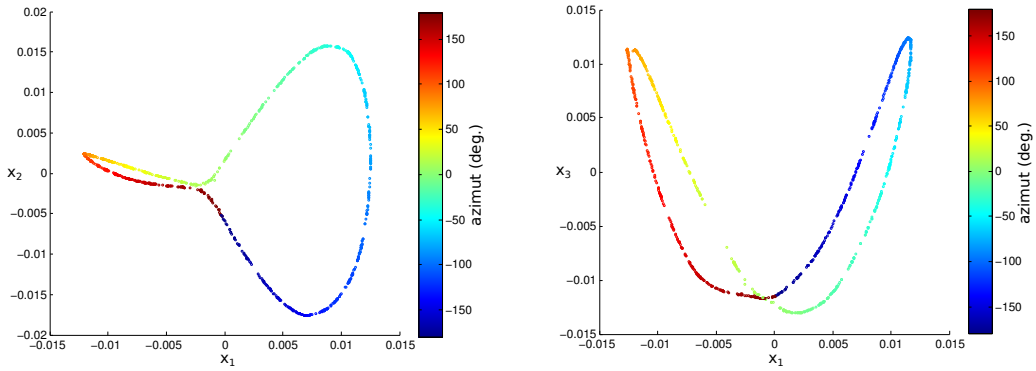
Chapitre 8

Ambiguïté perceptuelle

Sommaire

L'aspect actif du modèle du chapitre précédent, concrétisé par le comportement d'orientation, permet l'apprentissage d'une représentation de l'espace auditif à partir de laquelle une localisation sans déplacement devient possible. Ce modèle est motivé notamment par des arguments biologiques puisque le comportement d'orientation, présent chez le nouveau-né, semble intervenir dans le processus d'apprentissage de la localisation de sources sonores (Muir *et al.* , 1989; Metta, 2000). Cependant le chapitre 2 a mis en évidence que d'autres stratégies de perception active sont mises en oeuvre par le système auditif dans le cadre de la localisation. Nous proposons ici une modélisation des mouvements de tête qu'un auditeur peut effectuer pour lever une incertitude dans une tâche de localisation (voir le paragraphe 2.1.3). Ces mouvements de tête intègrent l'action à un plus haut niveau que le comportement d'orientation. En effet ce dernier est utilisé pour *construire* une représentation de l'espace sensoriel tandis que les mouvements de tête considérés ici permettent d'*explorer* une représentation de l'espace sensoriel déjà construite, dans le but de réduire l'erreur de localisation induite par un état sensoriel ambigu.

Ainsi ce chapitre propose un modèle sensorimoteur basé sur le concept d'ambiguïté perceptuelle qui permet à l'agent d'une part de quantifier le degré de confiance à accorder à une estimation de localisation et d'autre part à pouvoir minimiser cette



(a) Projection sur les axes (x_1, x_2).

(b) Projection sur les axes (x_1, x_3).

FIGURE 8.1 – Représentation 3D d'une variété sensorielle de l'espace azimutal présentant une ambiguïté avant/arrière. L'ambiguïté est matérialisée par la singularité proche de $(0, -0.01)$, où des points très proches dans l'espace sensoriel sont associés à des états moteurs distants de 180°. La variété, composée de 1000 points, est apprise dans les mêmes conditions que Fig. 5.5(c) (données CAMIL, ILD). La représentation est obtenue par l'algorithme des cartes propres Laplaciennes (voir l'annexe C.1.2). (a) Projection sur les axes (x_1, x_2). (b) Projection sur les axes (x_1, x_3).

ambiguïté de manière active, via une exploration de son espace sensorimoteur. Après avoir établi nos motivations, ce chapitre introduit la notion d'ambiguïté perceptuelle puis propose une stratégie d'exploration permettant de minimiser cette ambiguïté. Enfin la section expérimentale présente deux simulations, la première sur un cas théorique modélisant l'ambiguïté avant/arrière, la seconde sur une tâche de localisation azimutale de sources sonores. Nous montrons dans ces deux cas que la minimisation de l'ambiguïté perceptuelle mène à une amélioration significative des performances de localisation.

8.1 Motivations

Ce paragraphe revient sur deux aspects introduits au chapitre 5 et qui motivent l'utilisation de l'ambiguïté perceptuelle dans une tâche de localisation de sources sonores. Nous revenons premièrement sur l'ambiguïté avant/arrière et sa résolution, puis de manière plus générale à la question de l'échantillonnage de l'espace sensoriel.

8.1.1 Ambiguïté avant/arrière

Revenons sur l'expérience proposée au paragraphe 5.3.2, dans laquelle il est montré que les indices spectraux induits par les HRTF sont nécessaires à l'apprentissage de l'ambiguïté avant/arrière de manière passive. La Fig. 5.5(c) présente une variété sensorielle obtenue à partir de 2000 états sensoriels. La Fig. 8.1 quant à elle présente une variété sensorielle obtenue dans les mêmes conditions que Fig. 5.5(c) mais avec 1000 points seulement. Alors qu'avec 2000 points l'ambiguïté avant/arrière est complètement résolue, 1000 points ne suffisent pas à lever cette ambiguïté. La variété sensorielle obtenue présente en effet une singularité : des points très proches dans l'espace sensoriel sont associés à des états moteurs distants de 180°.

Comment se résout alors cette ambiguïté au cours de l'apprentissage ? Cela dépend en fait de la densité de l'échantillonnage au voisinage de ce point singulier : si cette densité est trop faible par rapport à l'ordre k du voisinage, la différence entre des points « avant » et des points « arrière », qui est faible du fait de l'ambiguïté, n'apparaît pas. À l'inverse à mesure que la densité locale augmente, l'estimation des k-ppv permet d'établir cette différence. Un sur-échantillonnage de la variété sensorielle apparaît donc une solution naturelle dans l'apprentissage de l'ambiguïté. Néanmoins le sur-échantillonnage uniforme tel que proposé Fig. 5.5(c) n'est pas le plus adapté, un sur-échantillonnage local au niveau de la singularité permettrai en effet de lever l'ambiguïté en exploitant un nombre minimal d'échantillons.

8.1.2 Échantillonnage de l'espace sensoriel

La topologie de l'espace sensoriel peut être complexe, ce qui rend son échantillonnage délicat. Aytekin *et al.* (2008) remarquent ainsi que la courbure est très importante pour des angles d'élévation élevés sur une variété auditive apprise à partir de HRTF, rendant difficile la localisation dans ces directions. Laflaquière (2013) montre qu'un étirement trop important de la variété sensorielle entraîne l'échec de l'estimation de sa dimension intrinsèque. Il propose également une méthode d'homogénéisation de la densité de l'échantillonnage permettant de résoudre ce problème.

Dans une approche similaire à ce travail, la notion d'ambiguïté perceptuelle présentée dans ce chapitre permet de repérer la « mauvaise qualité » d'une variété sensorielle, non pas du point de vue de sa topologie comme proposé par Laflaquière (2013), mais à partir de sa relation avec la variété motrice. Reprenons le cas des élévations élevées identifié par Aytekin *et al.* (2008) et supposons que chaque échantillon de l'espace sensoriel soit associé à un échantillon de l'espace moteur selon notre définition de la localisation (Eq. 5.3). Dans cette zone de forte élévation, deux points voisins dans l'espace sensoriel peuvent être associés à deux sources sonores aux positions éloignées et donc à points distants dans l'espace moteur. Dans ce cas de figure notre méthode de localisation par interpolation échoue et, s'il n'est pas possible du point de vue de l'agent, d'accéder à cette erreur de localisation, l'ambiguïté perceptuelle permet précisément d'identifier ces associations sensorimotrices problématiques.

8.2 Formalisation

Ce paragraphe formalise la notion d'ambiguïté perceptuelle puis introduit une stratégie de minimisation active de cette ambiguïté. Nous établirons au paragraphe suivant un lien entre cette ambiguïté et l'erreur de localisation de manière expérimentale. Les notations utilisées ici reposent sur le formalisme sensorimoteur introduit au paragraphe 5.2.

8.2.1 Ambiguïté perceptuelle

Soient un espace sensoriel $\mathcal{S}$, un espace moteur $\mathcal{M}$ et un espace environnemental $\mathcal{E}$. Soient également $S \subset \mathcal{S}^n$ et $M \subset \mathcal{M}^n$ des échantillonnages de $\mathcal{S}$ et de $\mathcal{M}$ respectivement, chacun de taille n et tels que pour tout $i \in [1, n]$, $m_i \in M$ correspond à l'estimation de localisation associée à s_i (voir Eq. 5.3). Soit enfin une carte sensorimotrice A, définie comme au paragraphe 5.2.3.2 associant chaque état sensoriel de S à son état moteur associé dans M, de sorte que pour tout $i \in [1, n]$ nous ayons $A = \{(s_i, m_i)\}$. Cette association peut par exemple être obtenue de manière

supervisée, à partir d'une base de donnée comme au chapitre 5, ou par apprentissage sensorimoteur comme au chapitre 6.

Considérant un état sensoriel $s \in \mathcal{S}$, son ambiguïté perceptuelle $\sigma(s)$ dans A est définie comme l'écart-type des états moteurs associés aux états sensoriels voisins de s dans S. Soient $K_S(s) = \{s_i | s_i \in S\}_{i\in[1,k]}$ l'ensemble des k voisins de s dans S et $K_M(s) = \{m_i | m_i \in M\}_{i\in[1,k]}$ les k états moteurs associés à $K_S(s)$, l'ambiguïté perceptuelle $\sigma(s)$ associée à s s'exprime comme :

$$\sigma(s) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(m_i - \overline{K_M(s)})^2}, \text{ avec } \overline{K_M(s)} = \frac{1}{k}\sum_{i=1}^{k} m_i \text{ et } m_i \in K_M(s). \quad (8.1)$$

Un état sensoriel ambigu est donc un état sensoriel dont le voisinage présente une forte variabilité de ses associations sensorimotrices. Ce cas se retrouve typiquement au niveau de la singularité présente Fig. 8.1, où nous avons vu que des états sensoriels très proches sont reliés à des états moteurs distants de 180°.

8.2.2 Minimisation active de l'ambiguïté

Supposons l'ambiguïté perceptuelle calculée pour tous les points de S. Supposons de plus un état sensoriel $s = \Phi(m, e)$ dont l'ambiguïté $\sigma(s)$ est significative. En adaptant l'état moteur m, il est alors possible d'explorer l'espace sensorimoteur à la recherche d'un état sensoriel le moins ambigu possible. Minimiser l'ambiguïté $\sigma(s)$ revient idéalement à estimer le déplacement moteur Δm tel que :

$$\Delta m = \underset{\tilde{m}\in\mathcal{M}}{\operatorname{argmin}}(\sigma(\Phi(m + \tilde{m}, e))). \quad (8.2)$$

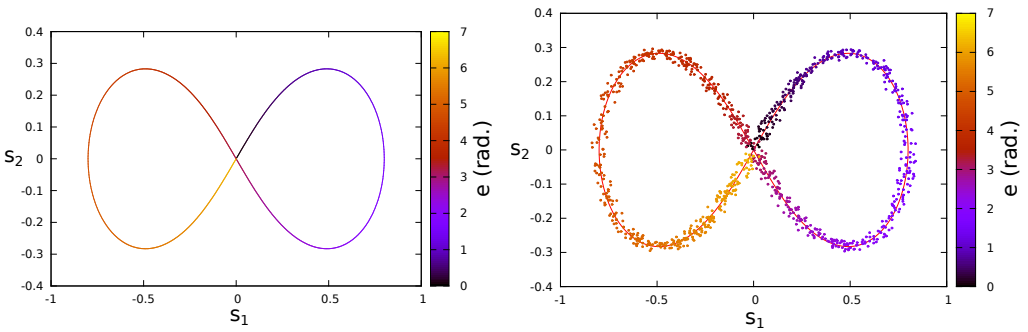
Cette équation consiste donc en la recherche d'un minimum global. Si une telle approche est théoriquement fondée, elle se révèle inapplicable en pratique puisqu'elle nécessite une exploration complète de $\mathcal{M}$. Nous proposons donc une stratégie d'exploration locale et itérative qui se veut la plus simple possible. En effet l'objectif n'est pas ici de proposer une stratégie d'exploration évoluée, il est de démontrer l'intérêt d'une telle stratégie dans une tâche de perception active.

En supposant l'environnement constant, la minimisation s'effectue par une exploration en « ligne droite » à partir de l'état moteur initial. A chaque itération l'état moteur est incrémenté par un δm constant, de sorte que $s_{i+1} = \Phi(m_i + \delta m, e)$. Cette opération est itérée jusqu'à atteindre une itération maximale it_{max} ou jusqu'à ce que l'ambiguïté courante tombe en dessous d'un seuil d'ambiguïté minimale σ_{min}. L'incrément moteur δm est calculé à l'initialisation comme un vecteur d'amplitude constante dont la direction est celle de l'état moteur m_{min} associé à l'état sensoriel s_{min} dont l'ambiguïté est minimale au voisinage de s (l'état sensoriel à la première itération) :

$$s_{min} = \underset{s_k\in K_S(s)}{\operatorname{argmin}}\,(\sigma(s_k)), \quad (8.3)$$

où $K_S(s)$ est l'ensemble des k voisins de s dans S. A chaque itération l'agent corrige donc son état moteur d'un faible incrément qui reste constant tout au long du processus.

Cette stratégie d'exploration nécessite l'introduction de deux paramètres : l'itération maximale it_{max} et le seuil d'ambiguïté minimale σ_{min}. Dans les simulations du paragraphe suivant ces paramètres seront fixés *a priori*, ce qui est difficilement



(a) Lemniscate de Bernouilli (b) Distribution d'apprentissage

Figure 8.2 – Génération de la distribution d'apprentissage à partir de la lemniscate de Bernouilli. (a) Lemniscate d'amplitude $a = 0.8$, avec m constant et $e \in [0, 2\pi]$ (dégradé de couleurs). Le point $(0, 0)$ associé à $m = 0(2\pi)$ (devant) et $m = \pi(2\pi)$ (derrière) constitue le point d'ambiguïté avant/arrière. (b) Distribution d'apprentissage composée de 1000 états sensoriels bruités issus de la lemniscate, associés à 1000 états environnementaux (la courbe rouge représente la lemniscate non-bruitée).

soutenable dans une approche sensorimotrice de la perception. Ce point sera discuté plus en détail au paragraphe 8.4.

8.3 Expériences

Ce paragraphe présente deux simulations. La première illustre la notion d'ambiguïté perceptuelle sur un exemple simple représentant un cas idéal d'ambiguïté avant/arrière. Le lien existant entre ambiguïté et erreur de localisation est montré, de même que la réduction de l'erreur de localisation par minimisation de l'ambiguïté. La seconde expérience reprend la même démarche mais appliquée cette fois-ci à une tâche de localisation azimutale de sources sonores. Une amélioration significative des performances de localisation est là encore démontrée.

8.3.1 Un cas idéal d'ambiguïté avant/arrière

Nous proposons ici une loi sensorimotrice reproduisant l'ambiguïté avant/arrière observée Fig. 5.5(c). Cette loi est exprimée de manière analytique par une lemniscate de Bernouilli. Après avoir décrit l'espace sensorimoteur et le protocole expérimental, nous montrons que la stratégie de minimisation de l'ambiguïté permet à l'agent de s'éloigner de la zone ambiguë et ainsi de réduire l'erreur de localisation.

8.3.1.1 La lemniscate de Bernouilli

La lemniscate de Bernouilli est une courbe ayant la forme du symbole ∞. Cette courbe, représentée Fig. 8.2(a), constitue une modélisation idéale de l'ambiguïté avant/arrière. La lemniscate inclut en effet une singularité en $(0, 0)$ représentant cette ambiguïté. Posons ainsi $\mathcal{S} = \mathbb{R}^2$ l'espace sensoriel défini par la lemniscate, $\mathcal{M} = [0, 2\pi]$ l'espace moteur et $\mathcal{E} = [0, 2\pi]$ l'espace environnemental. Soient $m \in \mathcal{M}$, $e \in \mathcal{E}$ et $(s_1, s_2) \in \mathcal{S}$, nous avons donc $(s_1, s_2) = \Phi(m, e)$. La loi sensorimotrice Φ est directement définie par l'équation paramétrique de la lemniscate. Elle est donc

donnée en fonction de $m \in \mathcal{M}$, de $e \in \mathcal{E}$ et d'une amplitude a comme :

$$\begin{cases} s_1 = & a\frac{u(1+u^2)}{1+u^4} \\ s_2 = & a\frac{u-u^3}{1+u^4} \end{cases} \text{, avec } u = \tan(\frac{e-m}{2}). \tag{8.4}$$

Nous avons $s_1 \in [-a, a]$ et $s_2 \in [-a/2\sqrt{2}, a/2\sqrt{2}]$ et nous posons pour la suite $a = 0.8$. Notons enfin que, de part la relation $e-m$, la compensation d'un changement environnemental s'effectue par un changement identique de l'état moteur visant à retrouver la perception environnementale initiale.

8.3.1.2 Protocole expérimental

La distribution d'apprentissage est composée de 1000 états sensoriels générés à partir de l'Eq. 8.4 pour 1000 états environnementaux aléatoires. Un bruit uniforme tiré dans l'intervalle $[-\alpha, -\alpha] \times [-\alpha/2\sqrt{2}, \alpha/2\sqrt{2}]$, avec $\alpha = 0.05a$, est de plus ajouté aux états sensoriels, comme illustré Fig. 8.2(b). De la même manière, une distribution de 1000 états sensoriels est générée pour les tests. Une fois les distributions générées, l'expérience se base en trois temps : (1) évaluation de l'erreur de localisation et de l'ambiguïté avant minimisation, (2) minimisation de l'ambiguïté des états sensoriels ambigus, et (3) évaluation de l'erreur de localisation et de l'ambiguïté après minimisation.

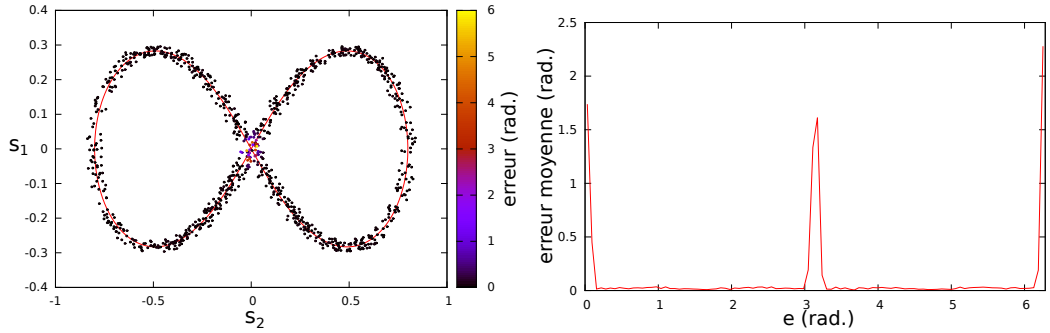
Pour chaque point s de la distribution de test, l'erreur de localisation est égale à $\Delta\tilde{m} = |\tilde{m} - e|$, où $\tilde{m}$ est donnée par interpolation au voisinage de s (Eq. 5.4 avec $k = 12$). Si l'ambiguïté $\sigma(s)$ associée à s est supérieure à σ_{min}, l'algorithme de minimisation de l'ambiguïté est exécuté et l'erreur de localisation après la minimisation est calculée comme précédemment. Les paramètres utilisés pour la minimisation sont $it_{max} = 100$, $|\delta m| = \frac{2\pi}{1000}$ et $\sigma_{min} = 0.1$. La valeur de $|\delta m|$ est choisie pour garantir un faible changement moteur à chaque itération, de sorte que l'état sensoriel minimisé ne soit pas complètement éloigné de la singularité. σ_{min} est choisi arbitrairement après analyse de l'ambiguïté sur l'ensemble de la distribution de test.

8.3.1.3 Résultats

Ce paragraphe détaille les résultats de localisation obtenus sur la lemniscate en termes d'erreur de localisation et d'ambiguïté perceptuelle, avant et après minimisation de l'ambiguïté.

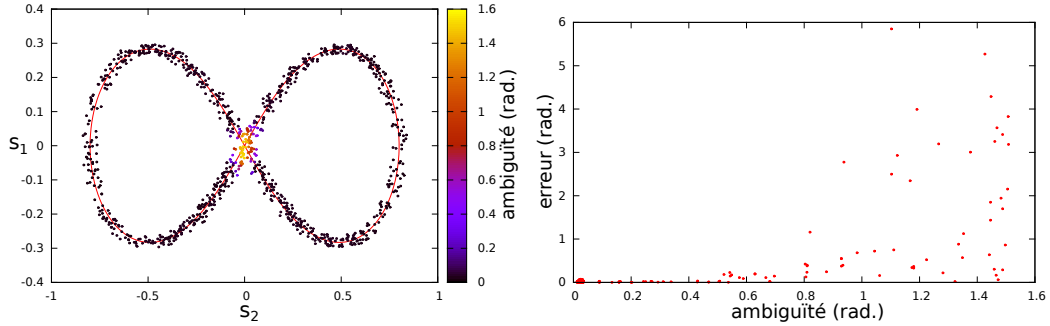
Avant minimisation L'erreur de localisation et l'ambiguïté perceptuelle obtenues sur la lemniscate sont présentées Fig. 8.3 et Fig. 8.4 respectivement. L'erreur de localisation se concentre sur les états sensoriels proches du point singulier (Fig. 8.3(a)), ce que confirme l'analyse de l'erreur moyenne en fonction de l'état environnemental (Fig. 8.3(b)). Ainsi l'erreur de localisation est proche de 0 sur l'ensemble de l'espace environnemental, à l'exception de $e = 0$, $e = \pi$ et $e = 2\pi$. Ce résultat démontre que l'ambiguïté avant/arrière est ici la seule source d'erreur, comme nous cherchions à le modéliser.

De la même manière que pour l'erreur, l'ambiguïté perceptuelle prend des valeurs élevées autour de la singularité, c'est-à-dire autour des états environnementaux $e = 0$, $e = \pi$ et $e = 2\pi$ (Fig. 8.4(a)). L'analyse de la répartition de cette ambiguïté perceptuelle obtenue en fonction de l'erreur de localisation (Fig. 8.4(b)) montre que



(a) Erreur de localisation (b) Histogramme de l'erreur

Figure 8.3 – Erreur de localisation avant minimisation. (a) Erreur de localisation sur la distribution de test. (b) Histogramme de l'erreur en fonction de l'état environnemental e.



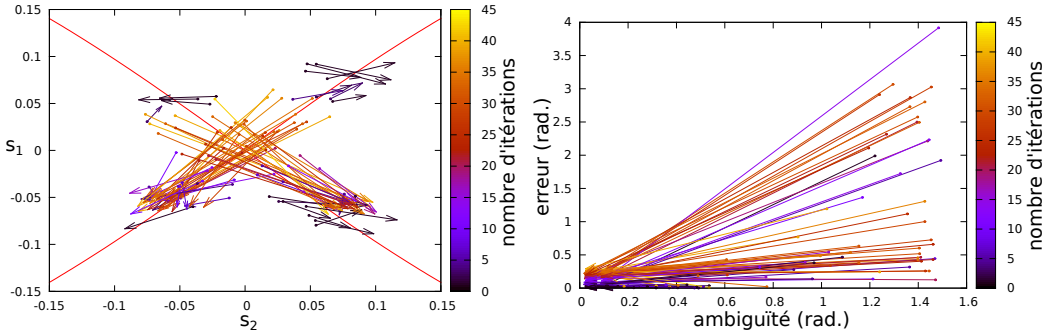
(a) Ambiguïté perceptuelle (b) Ambiguïté en fonction de l'erreur de localisation

Figure 8.4 – Ambiguïté perceptuelle avant minimisation. (a) Ambiguïté perceptuelle sur la distribution de test. (b) Ambiguïté en fonction de l'erreur de localisation.

tous les états sensoriels engendrant une erreur de localisation significative sont associés à ambiguïté importante. Notons cependant que la réciproque n'est pas vérifiée : certains états sensoriels ambigus engendrent une erreur de localisation très faible.

Après minimisation Le déroulement de la minimisation sur les échantillons de test ambigus est détaillé Fig. 8.5. Sur la lemniscate tout d'abord (Fig. 8.5(a)), il est constaté que les états sensoriels ambigus sont déplacés depuis le voisinage de la singularité vers sa périphérie. Cependant les expériences dont la minimisation a demandé le plus d'itérations (en orange et jaune Fig 8.5(a)) se sont effectuées dans la mauvaise direction. Dans ces cas de figure l'état sensoriel s'approche puis traverse la singularité, au lieu de s'en éloigner. Cet effet est dû au bruit qui, ajouté lors de la génération des données sensorielles, peut fausser l'estimation de la direction de δm (voir paragraphe 8.2.2). Ce problème pourrait être résolu par une stratégie d'exploration plus évoluée, actualisant par exemple la direction de l'incrément moteur à chaque itération.

L'évolution de l'erreur de localisation en fonction de l'ambiguïté perceptuelle (Fig. 8.5(b)) confirme là encore nos attentes. Il est en effet montré que la minimisation de l'ambiguïté entraîne dans tous les cas une diminution de l'erreur de localisation.



(a) Minimisation de l'ambiguïté sur la lemniscate

(b) Minimisation de l'erreur de localisation

FIGURE 8.5 – Minimisation de l'ambiguïté et de l'erreur de localisation. (a) Déplacement des états sensoriels ambigus pendant la minimisation (zoom autour de la singularité de la lemniscate dont la courbe est représentée en rouge). (b) Évolution de l'erreur de localisation avant et après minimisation de l'ambiguïté perceptuelle. Sur chacune des 2 figures chaque flèche représente la minimisation d'un état sensoriel : l'origine représente l'état initial, la terminaison représente l'état final et la couleur le nombre d'itération exécutées durant la minimisation.

En moyenne sur la distribution de test l'ambiguïté perceptuelle passe de 55° à 1.8° durant la minimisation. L'erreur de localisation moyenne passe quant à elle de 60° à 6.2°. Cette expérience a donc permis d'une part d'illustrer le concept d'ambiguïté perceptuelle sur un cas concret et d'autre part d'établir un lien entre minimisation de l'ambiguïté et minimisation de l'erreur de localisation, lorsque l'erreur est provoquée par une singularité sur la variété sensorielle.

8.3.2 Localisation azimutale

L'objectif de cette seconde expérience est d'appliquer l'algorithme de minimisation de l'ambiguïté perceptuelle sur un cas de localisation de sources sonores dans le plan azimutal. Après avoir présenté le protocole expérimental, ce paragraphe détaille les résultats obtenus en termes de performance de localisation, avant et après minimisation de l'ambiguïté. Il est montré que cette minimisation permet de réduire significativement l'erreur de localisation.

8.3.2.1 Protocole expérimental

Cette expérience est effectuée en simulation dans un milieu anéchoïque. La source sonore est un bruit blanc disposé à élévation nulle et distance fixe, l'azimut étant choisi aléatoirement dans l'intervalle $[-180°, 180°]$. Le signal binaural est ensuite obtenu par interpolation des HRTF (paragraphe 4.1.1.3). L'ILD est finalement calculée selon l'Eq. 4.14, avec les paramètres de la table 5.1. Dans cette simulation l'espace sensoriel est composé de l'ensemble des indices d'ILD possibles, donc $\mathcal{S} = [-1, 1]^{30}$. L'espace environnemental est défini par les azimuts possibles de la source, on a $\mathcal{E} = [-180°, 180°]$. L'espace moteur est défini par les différents angles d'orientation du cou dans le plan horizontal. On a donc $\mathcal{M} = [-180°, 180°]$. Finalement les distributions d'apprentissage et de test, constituées de 200 points chacune, sont générées pour des environnements aléatoires.

À partir de ces deux distributions, l'expérience est similaire à la précédente. Pour chaque point s de la distribution de test, l'erreur de localisation est calculée, de même que l'ambiguïté perceptuelle. Les points dont l'ambiguïté est supérieure à σ_{min} sont ensuite traités par l'algorithme de minimisation. Les paramètres utilisés sont $it_{max} = 100$, $|\delta m| = 1°$ et $\sigma_{min} = 3$. σ_{min} est choisi arbitrairement après analyse de l'ambiguïté sur l'ensemble de la distribution de test.

8.3.2.2 Résultats

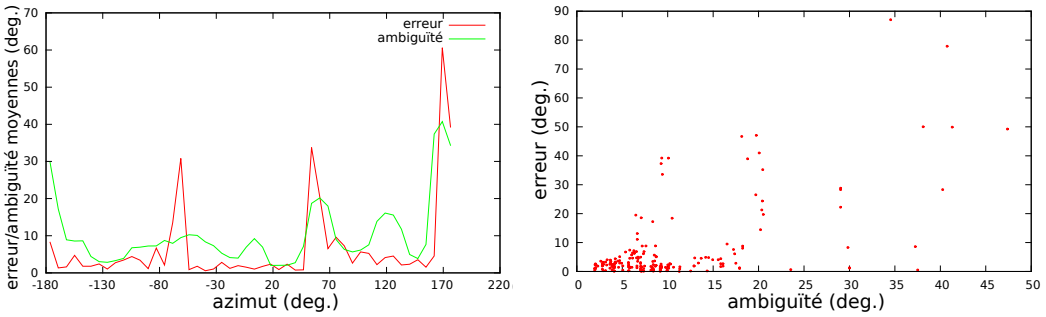
Ce paragraphe détaille les résultats de localisation obtenus dans cette tâche de localisation en termes d'erreur de localisation et d'ambiguïté perceptuelle, avant et après minimisation de l'ambiguïté.

Avant minimisation L'erreur de localisation et l'ambiguïté perceptuelle obtenues sur les 200 points de la distribution de test sont présentées Fig. 8.6(a). L'ambiguïté moyenne est à 10° tandis que l'erreur moyenne est à 7.2°. Il est constaté que l'ambiguïté avant/arrière n'est pas retrouvée, puisque l'erreur se concentre sur 3 azimuts distincts (autour de ±60° et de 170°). Ces fortes erreurs sont associées à une forte ambiguïté, à l'exception de la zone des −60°.

L'erreur de localisation en fonction de l'ambiguïté est présentée Fig. 8.6(b). Comme dans le cas de la lemniscate il est montré que tous les points ayant une erreur importante ont également une ambiguïté importante. Ainsi on peut s'attendre à ce que cette fois encore la minimisation de l'ambiguïté entraîne une réduction de l'erreur de localisation moyenne. Ces données justifient de plus le choix de $\sigma_{min} = 3°$. Notons enfin que nous aurions pu diminuer l'erreur moyenne simplement en augmentant le nombre de points dans la distribution d'apprentissage, ce qui n'est pas notre objectif. L'erreur est ici réduite par la perception active plutôt que par un meilleur échantillonnage de l'espace sensorimoteur.

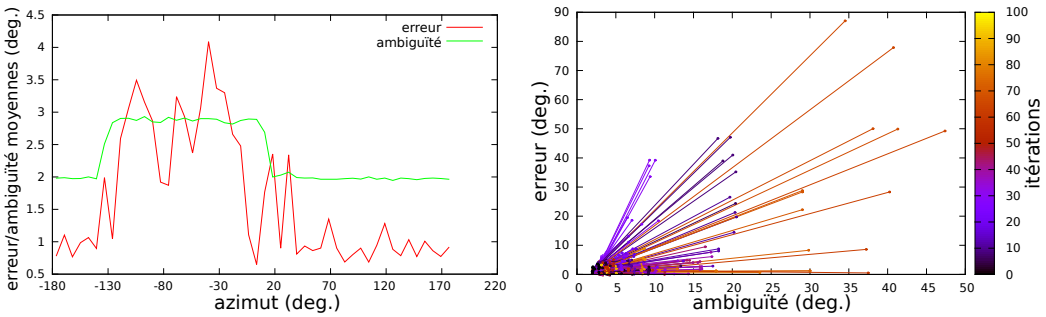
Après minimisation L'erreur de localisation et l'ambiguïté perceptuelle obtenues après minimisation sont présentés Fig. 8.7(a). Durant la minimisation, l'erreur de localisation moyenne passe de 7.2° à 1.6° et l'ambiguïté moyenne de 10° à 2.3°. Il est constaté que les 3 pics d'erreurs présents Fig. 8.6(a) ont été éliminés par la minimisation. L'erreur atteint en effet des valeurs autour de 1° sauf dans une large bande autour de −60° où elle varie autour des 3°. Cette erreur plus importante est également associée à la plus grande ambiguïté (3° environ, contre 2° ailleurs). La présence de cette bande de moindre qualité s'explique par la faible corrélation entre erreur et ambiguïté observée sur la Fig. 8.6(a) autour des −60°.

La Fig. 8.7(b) présente l'erreur de localisation en fonction de l'ambiguïté perceptuelle avant et après minimisation. Là encore les résultats sont comparables à ceux observés sur la lemniscate (Fig. 8.4(b)) : la minimisation de l'ambiguïté entraîne une réduction de l'erreur de localisation. Notons finalement le nombre généralement élevé d'itérations exploitées par le processus de minimisation. En effet chaque itération correspond à un déplacement de 1°, le déplacement total est donc bien plus important que celui typiquement observé chez l'humain, de quelques degrés seulement. Le déplacement total effectué par la minimisation est directement dépendant du seuil σ_{min} et, dans une moindre mesure, par la densité d'échantillonage de la distribution d'apprentissage. Là encore une stratégie d'exploration plus évoluée pourrait permettre de réduire le déplacement total.



(a) Erreur et ambiguïté moyennes (b) Ambiguïté en fonction de l'erreur

FIGURE 8.6 – Erreur de localisation et ambiguïté perceptuelle avant minimisation pour la distribution de test. (a) Histogrammes de l'erreur moyenne et de l'ambiguïté moyenne en fonction de l'azimut. (b) Erreur de localisation en fonction de l'ambiguïté.



(a) Erreur et ambiguïté moyennes (b) Erreur et ambiguïté moyennes

FIGURE 8.7 – Minimisation de l'ambiguïté et de l'erreur de localisation sur la distribution de test. (a) Histogrammes de l'erreur moyenne et de l'ambiguïté moyenne en fonction de l'azimut après minimisation. (b) Évolution de l'erreur de localisation pendant la minimisation de l'ambiguïté. Chaque flèche représente la minimisation d'un état sensoriel ambigu : l'origine représente l'état initial, la terminaison représente l'état final et la couleur le nombre d'itération exécutées durant la minimisation.

8.4 Discussion

Ce chapitre a montré au travers de deux simulations que la minimisation de l'ambiguïté perceptuelle permet de réduire significativement l'erreur de localisation. Il reste néanmoins à valider cette méthode dans un cadre plus réaliste, au travers une expérimentation sur plateforme robotique. De plus la stratégie d'exploration « en ligne droite » proposée ici pourrait être améliorée. L'incrément moteur par exemple peut être ré-estimé à chaque itération, de même que le seuil d'ambiguïté minimale peut être fixé par apprentissage. On peut également penser à des stratégies d'exploration non-itératives, basées par exemple sur le calcul d'une trajectoire de minimisation dans l'espace moteur. Ainsi si ce dernier chapitre ne propose pas d'expérience suffisamment complète pour juger de la pertinence d'une telle approche dans un cadre robotique, il permet encore une fois de démonter l'intérêt de la théorie sensorimotrice pour la modélisation et la conception de stratégies de perception active.

De plus cette approche considère l'action non plus pour l'échantillonnage de l'es-

pace sensoriel comme au chapitre 5 mais pour son exploration. Ainsi un modèle qui intégrerait apprentissage autonome et minimisation de l'ambiguïté combinerait ainsi perception active de bas-niveau (comportement d'orientation) et haut-niveau (minimisation de l'ambiguïté) dans un cadre unique. Un tel modèle permettrait un apprentissage adaptatif de l'espace sensorimoteur. D'une manière similaire à l'homogénéisation de distribution proposée par Laflaquière (2013), il serait en effet possible de privilégier l'apprentissage de nouvelles expériences dans les zones ambiguës en générant des actions volontaires augmentant localement la densité de l'échantillonnage.

Chapitre 9

Conclusion

Sommaire

L'objet de cette thèse était d'appliquer la théorie des contingences sensorimotrices à la localisation de sources sonores en robotique. Cette conclusion revient tout d'abord sur les principales contributions apportées par nos travaux, elle en détaille ensuite certaines limitations qui permettent de tracer quelques perspectives pour des recherches futures.

9.1 Contributions

Ce paragraphe revient sur les conclusions de chacun des chapitres et résume ainsi les contributions principales apportées par cette thèse.

Modèle binaural bioinspiré Le système auditif artificiel proposé au chapitre 4 permet d'extraire du signal binaural capturé par le robot les indices d'ILD et d'ITD nécessaires à la localisation. Une paire de modèles de cochlée filtre le signal d'entrée et le décompose en différentes plages fréquentielles, qui sont alors représentées sous forme d'impulsions, à la manière de l'information transitant sur le nerf auditif. L'ILD est alors calculé à partir des caractéristiques d'amplitude du train d'impulsions tandis que l'ITD se focalise sur ses caractéristiques temporelles.

Bien que ce modèle binaural comporte des éléments originaux, la représentation impulsionnelle et l'extraction des fronts d'onde notamment, l'objectif premier ayant guidé sa conception n'était pas d'apporter une amélioration par rapport à l'état de l'art. Il était plutôt de proposer un modèle intégré de localisation binaurale qui soit suffisamment robuste, efficace et simple d'utilisation pour une application robotique. L'audition binaurale en robotique est en effet un domaine de recherche encore récent et, à l'heure actuelle, il n'existe pas de « modèle standard » ou d'implémentation faisant référence. Le modèle présenté dans cette thèse propose ainsi une base solide pour l'investigation de la localisation sensorimotrice de sources sonores.

Approche sensorimotrice de la localisation L'approche sensorimotrice suivie dans cette thèse a été motivée par deux raisons. Premièrement l'étude des principes

biologiques de la localisation, développée au chapitre 2, a montré que cette tâche est fondamentalement active. De nombreux processus moteurs et efférents sont en effet impliqués dans la localisation, comme les mouvements des oreilles ou de la tête, le contrôle efférent de la cochlée, ou encore le système attentionnel.

Deuxièmement un état de l'art des méthodes de localisation en robotique, proposé au chapitre 3, a montré que ces modèles héritent bien souvent d'une conception passive de la perception : si une commande motrice est générée, elle est le fruit d'une succession de traitements purement passifs. Ces méthodes impliquent un effort de modélisation de la part du roboticien. Estimer un angle dans un espace physique à partir d'un signal donné nécessite en effet des connaissances à la fois sur le robot et son environnement, et lorsque la configuration robot-environnement se complexifie, ce travail de modélisation peut devenir problématique.

Perception active et intégration sensorimotrice En plaçant l'action au coeur du processus de localisation, l'approche sensorimotrice constitue un changement de paradigme par rapport à ces approches classiques. Les chapitres 5 et suivants ont proposé une définition sensorimotrice de la localisation, puis ont développé plusieurs modèles de localisation. L'exploitation des capacités motrices de l'agent a permi de ne reposer sur aucune connaissance extérieure, par opposition aux méthodes passives qui procèdent par modélisation. Premièrement les comportements d'orientation et de déplacement ont permis une forme primitive de localisation, indissociable du déplacement de l'agent. Le comportement d'orientation fut utilisé dans un second temps, à la fois pour l'échantillonnage de l'espace sensorimoteur et l'estimation d'un état sensoriel de référence qui correspond à la sensation d'une source « localisée ». L'échantillonnage sensorimoteur permet alors une localisation sans déplacement tandis que la connaissance de l'état de référence donne accès à une capacité d'auto-supervision.

Cette auto-supervision peut être considérée comme une forme élémentaire d'introspection, l'agent étant en mesure de juger de la qualité d'une estimation de localisation qu'il vient d'effectuer. Cette capacité d'introspection a été étendue par l'introduction du concept d'ambiguïté perceptuelle, qui permet à l'agent de quantifier le degré de confiance à accorder la sensation qu'il expérimente. Nous avons alors introduit une stratégie d'exploration de l'espace sensorimoteur permettant de minimiser cette ambiguïté de manière active. Il a été montré que cette minimisation entraîne une amélioration significative des performances de localisation.

Retour sur la localisation de sources sonores L'application de nos modèles de localisation a également permis de proposer des solutions originales à des problèmes classiques dans le domaine de la localisation binaurale. Ainsi le comportement d'orientation et de déplacement est en mesure de résoudre l'ambiguïté avant/arrière de manière totalement autonome, pour peu que le robot soit équipé d'oreilles externes suffisamment directives.

La localisation à partir d'un échantillonnage de l'espace sensorimoteur à également abouti à des résultats intéressants et non-triviaux, comme la résolution de l'ambiguïté avant/arrière de manière passive ou la localisation en élévation à partir d'indices binauraux usuellement utilisés pour la localisation azimutale. Il a été montré que ce sont les indices spectraux apportés par le filtrage des oreilles externes qui permettent ces capacités étendues de localisation. Cependant, et c'est ce qui rend ces résultats intéressants, aucun traitement spécifique n'est appliqué pour extraire ou interpréter ces indices spectraux. Ceux-ci déforment en fait la topologie de la variété

sensorielle et sont alors « capturés » par le processus d'échantillonnage sensorimoteur.

9.2 Limitations et perspectives

Ce dernier paragraphe revient sur les limitations de nos modèles et sur des sujets qui n'ont pas été abordés dans cette thèse et qui mériteraient un approfondissement. Cela permettra de proposer quelques perspectives pour des recherches futures.

Exploration de l'espace sensorimoteur La méthode d'apprentissage autonome de la localisation proposée au chapitre 7 se limite à la localisation azimutale. Cette limitation est causée par la dépendance de la méthode au comportement d'orientation. Étendre ce modèle à la localisation en élévation nécessiterait par exemple de développer un comportement d'orientation adapté. Une alternative plus intéressante serait d'exploiter l'état sensoriel de référence appris simultanément à la localisation azimutale. Il est en effet possible d'effectuer des déplacements moteurs autour de ce point de référence et d'en analyser les conséquences sensorielles. Des rotations du cou en élévation permettraient ainsi l'échantillonnage de l'espace sensorimoteur dans des zones encore inexplorées par l'agent puisque inaccessible au comportement d'orientation.

Le chapitre 8 a également montré que la minimisation active de l'ambiguïté perceptuelle pouvait aboutir à un apprentissage adaptatif de l'espace sensorimoteur, l'agent pouvant de manière autonome augmenter la densité de l'échantillonnage dans les zones localement ambiguës. Ces deux exemples de la localisation en élévation et de la minimisation de l'ambiguïté perceptuelle démontrent l'intérêt pour l'agent de disposer d'une stratégie d'exploration de l'espace sensorimoteur. Ceci constitue donc une voie de recherches pertinente pour une extension de notre modèle à des capacités de perception plus étendues.

Environnements complexes Les conditions expérimentales utilisées pour évaluer nos modèles de localisation sont restés simples : une source sonore unique, un spectre large-bande, un signal stationnaire, peu ou pas de bruit. Puisque l'approche sensorimotrice suivie dans cette thèse s'écarte de l'état de l'art, il était cependant nécessaire d'évaluer les capacités de localisation dans ces conditions basiques.

De plus la plupart des expérimentations ont eu lieu en simulation, que ce soit à partir de signaux artificiels ou d'enregistrements binauraux. Bien que les résultats présentés confirme la validité de notre approche dans le principe, une série supplémentaire d'expériences sur plateforme robotique est cependant nécessaire pour valider complètement l'applicabilité de notre approche à la robotique. Il serait également intéressant de procéder à des expériences de localisation multisources, permettant de quantifier la robustesse du modèle à une source distractrice par exemple.

Une expérimentation dans des environnements plus complexes et plus variés serait également pertinente. Cela permettrait en effet de quantifier le potentiel de généralisation du modèle à des environnements inconnus, ce qui est d'un intérêt évident pour la robotique autonome. Il s'agit par exemple de passer d'un environnement de bureau à un environnement urbain, d'une chambre sourde à un milieu réverbérant, ou encore d'un bruit blanc à des signaux de parole. Ces situations mènent en effet aux questions suivantes : l'agent a appris à localiser dans un contexte A, est-il en mesure de le faire dans un contexte B ? Comment évolue les performances de localisation dans A une fois que le contexte B est appris ? L'agent peut-il différencier ces

contextes, les comparer ? Cette capacité à discriminer des environnements différents peut sembler triviale mais elle mène directement à la perception de caractéristiques non-spatiales, telles que le niveau de bruit ambiant ou le niveau de réverbération.

Système efférent La chapitre 4 a introduit un certain nombre de paramètres permettant de moduler la réponse du modèle binaural, comme le seuil de transduction permettant l'inhibition des composantes de basse intensité du signal ou les temps d'intégration des indices binauraux permettant de stabiliser des signaux non-stationnaires. Bien que dans cette thèse la valeur de ces paramètres a été fixée *a priori*, ceux-ci sont modifiables dynamiquement et peuvent ainsi être contrôlés directement par l'agent. Nous avons qualifié ces connexions *top-down* de connexions efférentes.

Dans cette thèse nous avons restreint le concept de perception active à l'aspect moteur. Il est cependant possible d'étendre ce concept aux processus efférents. Supposons ainsi que les changements perceptifs induits par un changement de paramètres efférents puisse être compensés par un changement de l'état de l'environnement. Comme le déplacement d'une source sonore peut être compensée par un déplacement du robot, une baisse de l'intensité d'une source sonore peut en effet être compensée par une régulation du gain de la cochlée, l'apparition d'une source secondaire peut être compensée par une inhibition des canaux cochléaires concernés. Sous cette hypothèse de compensabilité des changements perceptifs induits par le système efférent, la théorie sensorimotrice semble ainsi offrir un formalisme tout à fait adapté au contrôle du système efférent, qui est alors assimilé au système moteur.

Modalités visuelle et tactile La nature multimodale de la perception est un aspect fondamental qui a été ignoré dans cette thèse. Nous nous sommes en effet focalisé sur la perception auditive et sur la localisation de sources sonores en particulier. Cependant l'approche sensorimotrice est par nature une approche amodale, l'espace sensoriel pouvant être constitué d'une modalité quelconque.

Il serait ainsi intéressant de transférer le modèle d'apprentissage de la localisation auditive à la modalité visuelle. Supposons l'agent équipé d'un capteur de luminance au centre duquel se trouve une fovéa. Le comportement d'orientation peut alors être envisagé comme un processus de maximisation de la luminance dans la zone fovéale. Du point de vue extérieur, l'apprentissage de l'espace visuel permet alors d'exprimer la position de chaque pixel de la caméra en termes moteurs. Il serait également intéressant d'envisager des traitements sensorimoteurs basés sur le flux optique.

Le sens du toucher est une modalité très intéressante, particulièrement dans le cadre de l'approche sensorimotrice. L'interaction que l'agent entretient avec un objet lorsqu'il le touche est en effet évidente : on appuie sur une éponge pour en déterminer la souplesse ou on frotte une surface du bout des doigts pour en percevoir la rugosité. La perception de la rugosité notamment, bien documentée chez l'humain et le rat, fait face à des ambiguïtés perceptuelles qui sont résolues par une variation des mouvements des doigts ou des vibrisses. Ainsi l'approche sensorimotrice semble là encore une approche pertinente pour la mise au point de systèmes de perception tactile.

Annexe A

Reproduction qualitative de l'effet de précédence

Cette annexe présente une simulation reproduisant de manière qualitative l'effet de précédence, une faculté du système auditif à ne pas percevoir les échos suffisamment brefs en condition réverbérante (voir paragraphe 2.1.2). Notre propos n'est bien sur pas de proposer un modèle de la perception humaine, cet effet de précédence apparaît en fait comme un « effet de bord » du modèle de Jeffress (voir paragraphes 4.2.1 et D.2).

Nous reproduisons ici en simulation une expérience classique permettant de rendre compte de l'effet de précédence chez l'humain. Le protocole consiste à diffuser simultanément deux sources au contenu identique mais à des positions différentes, la seconde, appelée *lag* ayant quelques ms de retard sur la première, appelée *lead*, comme illustré par la Fig. A.1. Ce protocole permet de créer de manière simple un écho artificiel parfaitement contrôlable. Ainsi, en faisant varier le retard Δt entre *lead* et *lag*, il est possible de caractériser les performances de localisation en fonction du temps de réverbération (Blauert, 1997; Litovsky *et al.* , 1999).

Le signal $s(t)$ utilisé ici est un bruit blanc normalisé et échantillonné à 44.1 kHz d'une durée de 1 s. A partir de ce signal initial sont générées les deux sources sonores, chacune située à 1 m de distance de l'auditeur. *Lead* est située à un azimut de $-45°$ et *lag* à un azimut de $45°$. Pour obtenir ces sources, le signal initial est convolué avec la HRTF correspondant aux deux angles utilisés (paragraphe 4.1.1, modèle 3). Les deux sources obtenues sont enfin diffusées avec un retard *lead/lag* variant de 0.01 ms

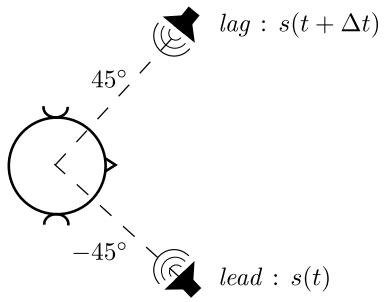


FIGURE A.1 – Configuration de la simulation. Deux sources *lead* et *lag* diffusent un signal identique $s(t)$, *lag* présentant par rapport à *lead* un retard Δt variant de 10^{-2} à 10^{2} ms.

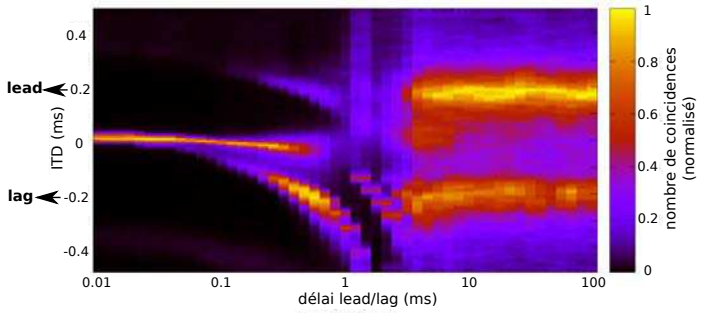


FIGURE A.2 – Reproduction de l'effet de précédence par le modèle de Jeffress. *Lead* et *lag* sont deux sources sonores identiques disposées à $-45°$ et à $+45°$, correspondant un ITD de 0.2 ms et -0.2 ms respectivement. Les sources sont diffusées avec un retard entre *lead* et *lag* variant de 0.01 ms à 100 ms et, pour chaque délai en abscisse, le signal d'ITD obtenu est normalisé.

à 100 ms selon une quantification logarithmique. La Fig. A.2 représente le résultat de cette simulation pour un modèle binaural utilisant une paire de cochlées à 30 canaux entre 100 Hz et 1000 Hz, un seuil d'extraction des pics fixé à 0 et une durée d'intégration de l'ITD T_{itd} à 0.1 s.

Comme nous le décrivons au paragraphe 2.1.2, l'effet de précédence se décompose, en fonction du délai *lead/lag*, en 3 phases successives : la fusion pour un retard très bref, la dominance pour un retard plus long, jusqu'à 35 ms, et la discrimination, pour des délai supérieurs à ce seuil d'écho. Ces 3 phases sont retrouvées qualitativement dans le *pattern* d'ITD présenté Fig. A.2. Ainsi pour un délai inférieur à environ 0.5 ms, les deux sources sont fusionnées en un percept unique dont la direction est centrée. Entre 0.5 ms et 5 ms environ, nous retrouvons la phase de dominance qui correspond en fait à un état transitoire. A l'inverse des expériences menées en psychoacoustique où la source *lead* domine la perception, c'est ici *lag* qui l'emporte. Finalement, pour des délais élevés, la discrimination s'opère et les sources *lead* et *lag* apparaissent comme séparée et localisée à leur véritable position. Ce résultat semble être une propriété intrinsèque de notre implémentation du modèle de Jeffress puisque nous n'utilisons ici ni seuil ni fronts d'ondes. Notons également que le seuil d'écho obtenu est moins important que celui observé en psychophysique (5 ms contre 35 ms). Des résultats similaires sont obtenus en remplaçant le bruit blanc en entrée par un train de clics, qui est le signal classiquement utilisé en psychoacoustique (Litovsky *et al.* , 1999). Nous retrouvons également les même résultats à partir d'une cochlée large bande (jusqu'à 8 kHz au lieu de 1 kHz dans cette simulation) mais, à cause des ambiguïtés provoquées par l'ITD dans les hautes fréquences, le *pattern* se trouve plus bruité.

Annexe B

Reconnaissance auditive et tactile de textures

Sommaire

Tandis que le reste de cette thèse a pour objet la localisation de sources sonores, cette annexe propose une modélisation de la perception de textures par les modalités auditive et tactile. La modalité tactile a en effet été modélisée sur le robot-rat Psikharpax, sous forme de matrices de vibrisses (N'Guyen *et al.* , 2011a) et les auteurs ont démontré les capacités de ce système à la reconnaissance de textures N'Guyen (2010).

Cette annexe reproduit ce modèle en simulation et l'étend à la modalité auditive. Il est montré que la discrimination de textures s'effectue avec de bonnes performances tant dans la modalité auditive que tactile et que les résultats obtenus sont similaires à ceux obtenus par N'Guyen (2010).

Les travaux présentés ici sont pour partie issus de Bernard *et al.* (2010b). La génération de textures artificielles est tout d'abord détaillée, un modèle de banc de vibrisses à base de gammatone est ensuite proposé, puis l'algorithme d'extraction des indices utilisés pour la reconnaissance est donné. Enfin le dernier paragraphe présente les résultas expérimentaux.

B.1 Génération de textures

D'une manière générale une texture peut s'assimiler à un signal stationnaire au spectre large. Concernant la modalité tactile, des jeux de papiers ponce aux grains différents sont ainsi généralement utilisés dans les expériences concernant la discrimination de texture, notamment par N'Guyen (2010). Neimark *et al.* (2003) modélisent un papier ponce dans le domaine fréquentiel comme un spectre à trois pics, avec une activité plus importante dans les basses fréquences. Conformément à ce modèle, nous avons généré un ensemble de 16 textures « virtuelles », 8 textures auditives et 8 textures tactiles, dont les fréquences des 3 pics, référencées Fig. B.1, sont choisies

Texture	F1	F2	F3
1	1	3	5
2	0.1	1	10
3	0.55	2	7
4	0.15	3	7
5	0.2	10	15
6	0.33	5	10
7	0.5	1	3
8	0.2	1	2

(a) Textures auditives. Fréquences en kHz.

Texture	F1	F2	F3
1	250	420	710
2	100	200	500
3	90	350	800
4	200	350	800
5	250	500	600
6	65	250	600
7	90	600	800
8	150	550	850

(b) Textures tactiles. Fréquences en Hz.

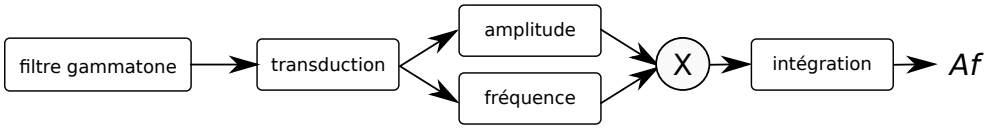
FIGURE B.1 – Composition spectrale des textures auditives (a) et tactiles (b). Sur la base d'un bruit blanc, F1 est amplifiée à 6 dB, F2 et F3 sont amplifiées à 3 dB, le reste du spectre est atténué à -3 dB.

arbitrairement. Le pic de basse fréquence est amplifié à 6 dB et les 2 autres à 3 dB. Le reste du spectre est atténué à -3 dB. Pour obtenir ces textures, nous avons généré un bruit blanc dont nous avons modifié le spectre à la main à l'aide du logiciel *Audacity*. Dans cette simulation chaque texture est donc un signal audio d'une durée de 40 s pour une fréquence d'échantillonage de 44.1 kHz pour les textures auditives et de 5 kHz pour les textures tactiles.

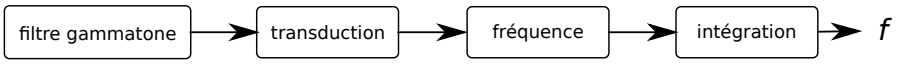
B.2 Modèle gammatone d'une matrice de vibrisses

Il existe de nombreuses similarités entre les modalités auditive et tactile pour la perception de textures, qui rendent l'étude de ce sujet tout à fait intéressante. L'audition et le toucher se basent en effet tous deux sur des traitements spectraux. Le système vibrissal du rat – ses moustaches – possède ainsi, comme la cochlée, une propriété de résonance permettant la décomposition spectrale du signal perçu en différentes bandes fréquentielles (Arabzadeh *et al.* , 2004, 2005; Andermann & Moore, 2008). Chaque vibrisse résonne en effet à une fréquence qui dépend principalement de sa taille et de son épaisseur.

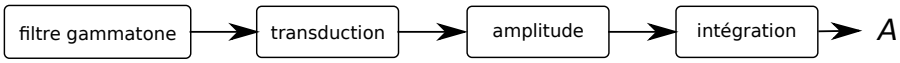
Les vibrisses du rat ont fait l'objet de plusieurs modélisations et implémentations, qu'elles soient computationnelles (Neimark *et al.* , 2003; Hartmann *et al.* , 2003) ou robotiques (Fend, 2005; Pearson *et al.* , 2007; Prescott *et al.* , 2009; N'Guyen *et al.* , 2011a). Nous proposons ici une approche originale puisque, à notre connaissance, les vibrisses n'ont jamais été modélisées à partir de filtres résonnateurs, en dépit de la ressemblance de la réponse impulsionnelle d'un filtre gammatone avec celle d'une vibrisse (Hartmann *et al.* , 2003). Cette approche évidemment se focalise sur la réponse fréquentielle de la perception tactile, en ignorant complètement la distribution spatiale des vibrisses le long du museau. Ainsi dans cette modélisation chaque vibrisse est représentée par un filtre gammatone dont les paramètres ont été adaptés. Neimark *et al.* (2003) ont mesuré les fréquences de résonance de 20 vibrisses de rats *in vivo*, décrivant 5 arcs de vibrisses composés chacun de 4 vibrisses. La fréquence de résonance des vibrisses est comprise entre 80 Hz et 750 Hz et, sur un arc donné, elles sont réparties des basses vers les hautes fréquences en allant de l'avant vers l'arrière du museau. Ces mesures *in vivo* sont utilisées pour fixer la fréquence



(a) Extraction de l'indice Af



(b) Extraction de l'indice f



(c) Extraction de l'indice A

Figure B.2 – Extraction des indices auditif et tactile pour la perception de textures à partir de la sortie d'un filtre gammatone. (a) modèle de cortex en tonneau, produit de l'amplitude A et de la fréquence f. (b) indice en fréquence f seulement. (c) indice en amplitude A seulement.

caractéristique des filtres gammatones de notre modèle, composé donc de 20 filtres. La bande passante à laquelle sont sensibles les vibrisses n'ont à notre connaissance pas été quantifiées et, faute de données disponibles, nous utilisons l'Eq. 4.6 et nous fixons la réponse asymptotique des filtres à $ear_Q = 35$ et la bande passante minimale $min_{BW} = 15$ (voir le paragraphe 4.1.2.2 pour une description de ces paramètres). Ces valeurs simulent une bande passante comprise entre 111 Hz pour les fréquences de résonances les plus basses à 233 Hz pour les plus élevées (moyenne à 145 Hz pour les 20 filtres).De même que le banc de filtres gammatone se limite à la modélisation de la réponse linéaire de la membrane basilaire, notre modèle de matrice de vibrisses possède de sérieuses limitations : l'organisation spatiale et bilatérale des vibrisses le long du museau, ou encore les aspects actifs du comportement de *whisking* ne sont ainsi pas modélisés. Lorsque le rat perçoit une texture du bout de ses vibrisses, de même que l'humain le ferait du bout des doigts, celui-ci effectue en effet des mouvements de va et vient, appelés mouvement de *whisking* dont la période est de 100 ms environ (Vincent, 1912). La modalité tactile est en effet une modalité active par nature et, sans ces mouvements, toute perception serait impossible.

B.3 Extraction d'indices pour la reconnaissance de textures

N'Guyen (2010) expérimente la reconnaissance de textures sur le robot-rat Psikharpax équipé de bancs de vibrisses en fibre de carbone fixée sur une peau artificielle en élastomère (N'Guyen *et al.* , 2009). Les performances de discrimination entre 8 papiers ponce de grains différents atteignent 90% en moyenne sur cette plateforme. Les auteurs utilisent pour cela un algorithme d'extraction d'indices inspiré du cortex en tonneau du rat, la partie de son cortex dédiée à la perception de textures. Un « tonneau » est ainsi une structure neuronale qui reçoit ses entrées d'une vibrisse principale, avec une faible influence des vibrisses voisines (Petersen, 2007). La recherche d'une base neurale à la représentation de texture au sein de cette structure a fait apparaître à Arabzadeh *et al.* (2004) une quantité homogène au produit Af

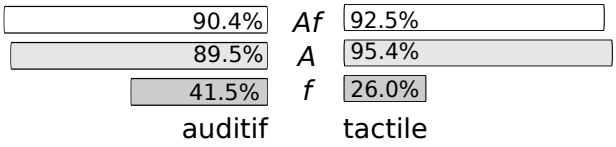
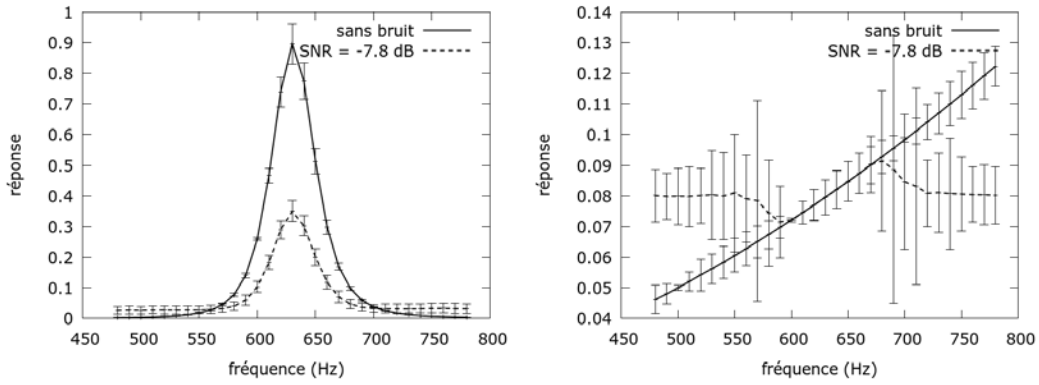


FIGURE B.3 – Taux de reconnaissance moyen des 8 textures obtenu pour les 3 indices Af, A et f sur les ensembles de test.



(a) Indice A. (b) Indice f.

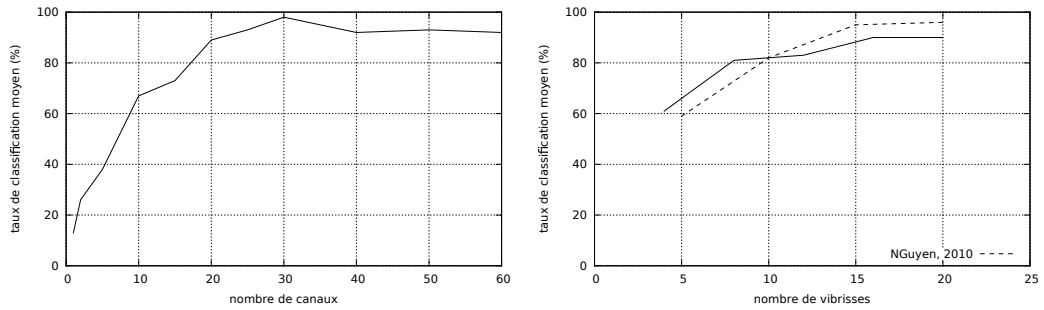
FIGURE B.4 – Comparaison des indices A et f calculés à partir de la vibrisse A4 (fréquence de résonance à 630 Hz) en réponse à un ton pur de fréquence variable autour de sa fréquence de résonance, sans bruit puis avec un bruit blanc important (SNR à -7.8 dB).

de l'amplitude A et de la fréquence f d'une sinusoïde stimulant les vibrisses.

L'algorithme utilisé ici, identique à celui de N'Guyen (2010), estime la fréquence instantanée f du signal d'entrée en calculant l'inverse de l'intervalle de temps entre deux impulsions détectées par le modèle de transduction, avec $\tau = 0$ (voir paragraphe 4.1.3). L'amplitude instantanée A est calculée de manière similaire à l'énergie (Eq. 4.13). La durée d'intégration, identique pour les deux modalités, est fixée à 100 ms ce qui correspond à la durée moyenne d'un mouvement de *whisking* du rat. Ainsi nous utilisons par la suite l'indice composé du produit Af mais nous étudierons également le comportement individuel de A et f (voir Fig. B.2). Notons que f est ici largement déterminé par la fréquence centrale du filtre auquel il est associé, ce qui en fait *a priori* un indice peu informatif. Nous le conservons néanmoins par soucis de comparaison avec les résultats de N'Guyen (2010). Par simplification nous désignons par la suite Af, A et$f$ comme des indices distincts.

B.4 Résultats expérimentaux

Pour chacun des 3 indices et chacune des 16 textures, un jeu de 400 vecteurs est généré (signal de 40 s pour une intégration à 100 ms). Le calcul des indices s'effectue en parallèle à la sortie de chaque filtre gammatone, comme illustré Fig. B.2 et la dimension des vecteurs est donc égale à 30 ou à 20 pour les modalités auditive ou tactile respectivement. Une fois ces vecteurs sensoriels générés, deux systèmes d'apprentissage supervisé sont mis en place, un par modalité. Pour chaque texture, 300 vecteurs sont utilisés pour l'apprentissage et 100 pour le test. Nous utilisons ainsi



(a) Modalité auditive. (b) Modalité tactile.

Figure B.5 – Influence du nombre de filtres sur le taux de classification. Taux moyen obtenu pour les 8 textures dans les deux modalités. (a) modalité auditive. (b) modalité tactile, en pointillé les résultats obtenus par N'Guyen (2010).

un perceptron multicouches (MLP) dans lequel la première et la seconde des 3 couches de neurones ont la même dimension que les vecteurs d'entrée, tandis que la couche de sortie contient 8 neurones, un par texture à discriminer. L'implémentation du MLP est fournie par la bibliothèque FANN (Nissen, 2003). Un algorithme d'apprentissage à base de rétropropagation est utilisé (Igel & Hüskel, 2000). Finalement la classification des textures s'effectue par la stratégie du *winner take all* sur la couche de sortie.

Les résultats de classification obtenus en sortie du MLP sont présentés Fig. B.3. La capacité de notre modèle à discriminer les différentes textures y est clairement démontrée pour les modalités auditive et tactile, le taux de classification moyen avoisinant les 90% pour les indices Af et A. Ceci confirme donc les résultats obtenus par N'Guyen (2010) sur plateforme robotique avec l'indice Af. Il est de plus montré que l'indice f considéré isolément ne permet pas de discriminer les textures entre elles : les résultats, 26.0% en auditif et 41.5% en tactile, sont au dessus d'une classification aléatoire (12.5% pour 8 textures) mais bien en deçà des performances atteintes par les indices à base d'amplitude. Nous constatons néanmoins que l'indice A atteint des performances très proches de Af, tandis que l'indice fréquentiel pur f se révèle inadapté du fait de sa trop étroite relation avec la fréquence de résonance des filtres.

Cette perte de performance liée à l'utilisation de l'indice fréquentiel se comprend en effet en observant la Fig. B.4, où est fournie une comparaison de la réponse des indices A et f pour une seule vibrisse (fréquence de résonance à 630 Hz), en réponse à un ton pur dont la fréquence varie autour la fréquence caractéristique de cette vibrisse. Deux conditions sont étudiées : sans bruit tout d'abord, puis avec un bruit important (SNR à -7.8 dB). Nous observons que l'indice A tire directement profit de la propriété de résonance du filtre gammatone, sa réponse étant significativement amplifiée autour de 630 Hz. De plus cet indice s'avère robuste à la présence de bruit. A l'inverse l'indice f ne permet pas une discrimination aisée de textures : si en l'absence de bruit, la réponse en sortie est linéaire à la fréquence en entrée, la variance importante peut entraîner des confusions. De plus le cas bruité nous révèle que l'indice f n'est pas robuste à ces conditions, et à l'exception d'une zone d'environ 50 Hz autour de la fréquence de résonance où les réponses dans les 2 conditions se superposent, l'estimation obtenue est constante et sa variance est très importante.

Une dernière expérience, présentée Fig. B.5 présente l'évolution du taux de classification moyen obtenu pour les 8 textures, à partir de l'indice Af uniquement, en

fonction du nombre de filtres présents en entrée. La même méthodologie que précédemment est employée. Les résultats nous montrent que le taux de classification converge rapidement avec l'augmentation du nombre de filtres, atteignant 80% environ à partir de 20 canaux pour la modalité auditive et 2 arcs (8 vibrisses) pour la modalité tactile. Ceci confirme des résultats précédemment obtenus concernant tant la reconnaissance tactile de textures (Fend *et al.* , 2003; N'Guyen, 2010) et que la localisation de source sonores (Deleforge & Horaud, 2011).

Annexe C

Localisation après réduction de dimension

Sommaire

Cette annexe se compose de deux parties. Tout d'abord sont introduits le principe de la réduction de dimension et l'algorithme des cartes propres Laplaciennes utilisé aux paragraphes 5.3.2 et 8.1.1. Enfin cette annexe propose une comparaison des performances de localisation dans l'espace sensoriel et sa représentation en basse dimension. Il est montré que la localisation après réduction de dimension n'apporte pas de gain de performance par rapport à une localisation directement dans l'espace sensoriel.

C.1 Réduction de dimension

Ce paragraphe revient sur la dimensionalité de l'espace sensoriel avant de présenter le principe de la réduction de dimension et son utilisation dans un contexte sensorimoteur. Enfin l'algorithme de s cartes prpores Laplaciennes utilisé dans cette thèse est détaillé.

C.1.1 Dimensionalité de l'espace sensoriel

L'espace sensoriel $\mathcal{S}$ est une variété dont la dimension intrinsèque peut être très inférieure à la dimension de l'espace dans lequel elle est plongée (Philipona *et al.* , 2003). Il est ainsi possible de trouver une paramétrisation de $\mathcal{S}$ reposant sur une dimensionalité plus petite par le biais de méthodes de réduction de dimension. Bien

qu'il ne s'agisse pas d'une nécéssité théorique, il existe au moins deux raisons pratiques à cette opération (Laflaquière, 2013). Premièrement, du point de vue extérieur, il est difficile d'appréhender et d'analyser des données en grande dimension et le passage dans un espace réduit permet une visualisation et une investigation plus aisée des données et du déroulement de l'apprentissage sensorimoteur. Dans un second temps, les calculs et la consommation mémoire induits par un nombre important d'échantillons représentés en grande dimension peuvent être handicapant pour une implémentation robotique, le passage en basse dimension permettant alors de répondre à ce problème. Il faut néanmoins relativiser cette solution car l'apprentissage d'une variété est lui aussi coûteux en temps de calcul, *a fortiori* en présence de nombreux points ou d'une dimension élevée de l'espace d'entrée.

Parmi les différentes méthodes de réduction de dimension existantes, la plus simple est l'analyse en composantes principales (PCA) (Philipona *et al.* , 2003). Cette méthode linéaire, qui consiste en une réduction des données selon les axes de plus forte variance, s'avère assez limitée en pratique, du fait de la potentielle non-linéarité de la loi sensorimotrice Φ et des variétés environnementale et motrice sous-jacentes (Couverture & Gas, 2009). De très nombreuses méthodes non-linéaires sont également proposées par la littérature (Van der Maaten *et al.* , 2009; Lee & Verleysen, 2010). Citons parmi celles-ci la LTSA utilisée notamment par Aytekin *et al.* (2008) et Deleforge & Horaud (2011), les SOM utilisées par exemple par Berglund *et al.* (2008) , ou encore l'analyse en composantes curvilignes (CCA) utilisée par Laflaquière (2013).

L'espace sensoriel $\mathcal{S}$ est donc une variété plongée dans un espace de grande dimension. L'opération de réduction de dimension permet d'obtenir une représentation en basse dimension de l'esapce sensoriel. Par simplification nous appellons cette représentation *espace de représentation*, qui est notée $\mathcal{R}$. Plus précisément, considérant un ensemble $S = [s_1, ..., s_n]$ de n vecteurs sensoriels correspondants à différents états moteurs et environnementaux, avec $S \in \mathcal{S}^n$, de sorte que pour chaque $i \in [1, ..., n]$ nous ayons $s_i = \phi(m_i, e_i)$. Nous pouvons exprimer l'échantillonnage $R = [r_1, ..., r_n]$ de l'espace de représentation $\mathcal{R}$ en fonction de l'échantillonnage S de $\mathcal{S}$ comme :

$$R = P(S), \tag{C.1}$$

où P représente la transformation opérée par l'algorithme de réduction de dimension. Nous avons donc $dim(R) < dim(S)$.

C.1.2 Cartes propres Laplaciennes

Nous utilisons ici l'algorithme des cartes propres Laplaciennes (LE) (Belkin & Niyogi, 2002, 2003; Baghani & Araabi, 2006), dans l'implémentation proposée par Van der Maaten *et al.* (2009). L'algorithme du LE est une méthode non-linéaire qui se propose d'estimer la variété de basse dimension inclue dans un espace de grande dimension sur la base de critères géométriques locaux, plus précisément en cherchant à minimiser les distances locales entre les points dans l'espace de sortie.

D'un point de vue formel, nous considérons l'espace d'entrée $S = [s_1, ..., s_n]$ composé de n points échantillonnant la variété sensorielle $\mathcal{S}$. L'objectif de LE est d'estimer une représentation basse dimension $R[r_1, ..., r_n]$ de l'espace sensoriel S. Pour ceci, l'algorithme minimise l'erreur de projection E_{LE} suivante :

$$E_{LE} = \frac{1}{2} \sum_{i,j=1}^{n} w_{i,j} ||r_i - r_j||^2, \tag{C.2}$$

où l'opérateur ||.|| représente la norme euclidienne et où les $w_{i,j}$ sont les éléments de la matrice symétrique W calculée à partir de la matrice d'adjacence $A = [a_{i,j}]_{i,j\in[1,n]}$ représentant le graphe de k-voisinage de S. Ainsi $w_{i,j} = 0$ si $a_{i,j} = 0$, c'est à dire si s_i et s_j ne sont pas voisins dans S, et $w_{i,j} \geq 0$ sinon. Belkin & Niyogi (2003) recommandent l'utilisation d'un noyau de chaleur (*heat kernel*) assimilable à une fonction gaussienne, de sorte que :

$$w_{i,j} = e^{-\frac{||s_i - s_j||^2}{2\sigma^2}}, \tag{C.3}$$

où σ peut être vu comme le paramètre de température du noyau. Ainsi, au travers de l'utilisation de W, la minimisation de E_{LE} nous assure que si s_i et s_j sont proches dans l'espace d'entrée (respectivement éloignés), leurs projections r_i et r_j seront proches elles aussi dans l'espace de sortie (respectivement éloignées), c'est à dire que la topologie locale de S sera respectée dans R. La minimisation de E_{LE} s'effectue de manière analytique par une décomposition en valeurs propres de la matrice laplacienne normalisée associée au graphe de voisinage de S (voir Lee & Verleysen (2010) pour plus de détails et une démonstration de ce raisonnement). Considérant ainsi l'espace R de dimension d_r, les vecteurs propres associés aux d_r plus petites valeurs propres calculées - exceptée la dernière qui est systématiquement nulle (Lee & Verleysen, 2010)) - forment les projections des s_i dans R.

Cet algorithme LE permet donc de conserver les caractéristiques locales de la variété tout en restant robuste au bruit et aux données aberrantes (Belkin & Niyogi, 2003). Il peut également être utilisé pour du *clustering* de données, des vecteurs propres étant alors estimés indépendamment sur chaque composante connexe du graphe. Le LE, dans l'implémentation proposée par Van der Maaten *et al.* (2009) que nous utilisons, est paramétré par 3 variables : la dimension d_r de la variété apprise que nous fixons à $d_r = 2$, l'ordre du voisinage k pour le calcul des k-ppv fixé à $k = 12$ et la variable σ paramétrant le noyau de chaleur, que nous laissons à sa valeur par défaut $\sigma = 0.1$.

Considérant enfin les indices d'ITD, ceux-ci sont normalisés dans l'intervalle $[0, 1]$ avant l'application de LE (les estimations se faisant directement dans S sont effectuées sant cette étape de normalisation). Ces vecteurs d'ITD sont en effet constitués à l'origine de valeurs entières (voir le paragraphe 4.2.1) qui font échouer l'alogorithme de réduction de dimension. Nous considérons pour la normalisation le maximum global des vecteurs d'apprentissage et de test.

C.1.3 Projection dans l'espace de représentation

Une fois la variété R apprise, nous avons besoin de pouvoir y projeter de nouveaux points, correspondant à de nouveaux états sensoriels encore inconnus du système. Nous aurons en effet besoin de cette propriété au paragraphes 5.3 et 7.1 quand nous chercherons à localiser de nouvelles sources sonores dans une variété précedemment apprise. Bien que LE soit une méthode non-générative, c'est à dire qu'elle ne permet pas d'estimer explicitement la transformation de S vers R, Bengio *et al.* (2003) proposent une extension aux LE permettant d'y projeter des points ne faisant pas partie de l'ensemble d'apprentissage. Considérant ainsi un nouveau point $\tilde{s} \in \mathcal{S}$ et l'estimation R de l'espace de représentation précédemment appris, la projection $\tilde{r}$ de $\tilde{s}$ dans R s'exprime grâce à la fonction de projection P_e comme :

$$\tilde{r} = P_e(\tilde{s}, R). \tag{C.4}$$

Appliqué à LE, cet algorithme se base sur la fonction de Nyström (Baker, 1977), originalement utilisée pour accélérer la décomposition en valeurs propres de matrices et qui repose sur le fait que l'estimation des paires valeurs propres/vecteurs propres converge avec l'augmentation du nombre de points composant la matrice (Bengio *et al.* , 2003). Nous ne rentrerons pas ici dans les détails de cette méthode basée sur la décomposition en valeurs propres d'une fonction noyau, ceux-ci étant proposés par Bengio *et al.* (2003). De même que l'alogorithme des LE, l'implémentation proposée par Van der Maaten *et al.* (2009) est utilisée.

C.2 Localisation comparée dans S et R

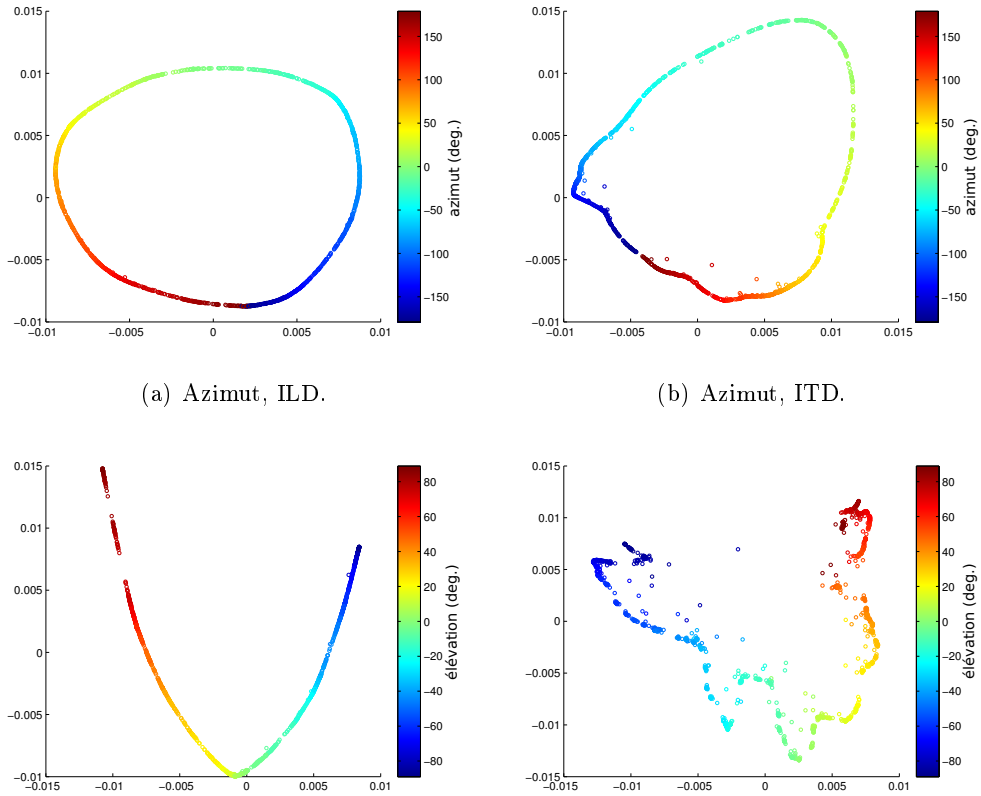
Cette expérience vise à comparer les performances de localisation obtenues dans l'espace sensoriel S et dans l'espace de représentation R. Le protocole utilisé ici est le même que celui proposé au paragraphe 5.3.1.1 et se base sur les enregistrements CAMIL. Les calculs sont effectués sur les indices d'ILD et d'ITD. Ce paragraphe détaille premièrement l'estimation R obtenue après application de LE. Les performances de localisation dans S et R sont ensuite comparées et il est montré que les résultats dans R sont légèrement en deçà de ceux obtenus dans S. Enfin, l'influence de la taille de S et de la dimension de R sur les performances de localisation est évaluée.

C.2.1 Apprentissage des variétés

A partir des vecteurs d'indices binauraux calculés comme détaillé au pargraphe 5.3.1.1, nous appliquons LE sur un sous-ensemble de 1500 vecteurs tirés aléatoirement, indépendamment en ILD et en ITD, et indépendamment en azimut et en élévation. Bien que la dimension réelle du problème soit 1 (un seul degré de liberté, soit en azimut soit en élévation), nous fixons la dimension d'apprentissage à $d_r = 2$. En effet, puisque la variété azimutale présente une boucle, un apprentissage en dimension 1 entraînerai une perte de sa topologie intrinsèque et le cercle obtenu serait alors brisé et transformé en courbe. Les 4 variétés obtenues après application de LE sont illustrées Fig. C.1. Ces résultats appellent deux commentaires.

Premièrement, que ce soit en considérant l'azimut ou l'élévation nous voyons que la topologie obtenue en ILD et en ITD est similaire (circulaire en azimut, parabolique en élévation). Les variétés obtenues en ITD sont néanmoins plus bruitées que leurs équivalentes en ILD, des points aberrants étant présents (nous proposerons au paragraphe 7.1 une méthode permettant de traiter ces points aberrants) , ainsi que des déformations locales relativement homogènes en élévation et concentrées derrière la tête en azimut. Cette différence de « qualité » entre les variétés d'ILD et d'ITD peut être causée par la plus forte dimensionnalité de l'espace des ITD, demandant ainsi plus de points pour parvenir à une qualité équivalente, mais une autre cause rentre également en jeu. Nous avons vu en effet au paragraphe 4.3 que l'indice d'ITD est plus sensible aux perturbations environnementales que l'indice d'ILD, rendant d'autant plus complexe l'apprentissage d'une représentation en basse dimension dans des conditions acoustiques réalistes.

Deuxièmement nous observons que l'apprentissage de l'élévation s'effectue convenablement à partir des indices binauraux théoriquement dédiés à la localisation en azimut uniquement. Ce constat peut surprendre car il est peut documenté par la littérature mais nous l'expliquons par le fait que les non-linéarités causées à la fois



(a) Azimut, ILD.

(b) Azimut, ITD.

(c) Élévation, ILD.

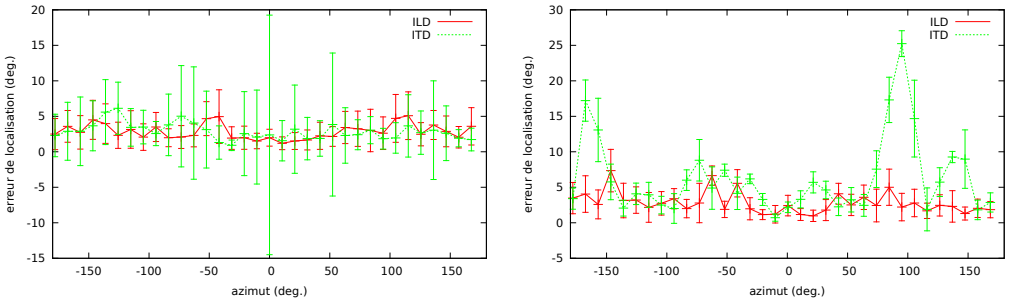
(d) Élévation, ITD.

Figure C.1 – Variétés auditives apprises par LE à partir des indices binauraux en azimut et en élévation à partir de la base CAMIL. Chaque variété est composée de 1500 points tirés au hasard parmi les points générées (3600 en azimut, 1800 en élévation). (a et c) variétés à base d'ILD. (b et d) variétés à base d'ITD. (a et b) variétés représentant 180 azimuts différents compris entre -180° et 180°, élévation constante à 1°. (c et d) variétés représentant 90 élévations différentes comprises entre $-90°$ et 90°, azimut constant à 1°.

par les HRTF et une éventuelle asymétrie des conditions acoustiques selon l'angle d'élévation (les réverbérations provenant du sol ou du plafond sont différentes) se reflètent dans les indices binauraux, comme nous l'avons observé dans l'apprentissage implicite des HRTF associé à la désambiguïté avant-arrière (voir Fig. 5.5). L'asymétrie des pavillons a également été rapportée par Searle *et al.* (1975) pour contribuer à la localisation en élévation à partir des indices binauraux mais la tête binaurale utilisée dans cette expérience ne semble pas présenter de telles asymétries (voir Fig. 3.1(c)).

C.2.2 Performances de localisation

Nous revenons ici sur le détail des performances de localisation obtenues à partir de l'interpolation aux k voisins présentée dans l'Eq. 5.4. A partir des ensembles de test générés en azimut et en élévation, pour l'ILD et l'ITD, nous sélectionnons de manière aléatoire 1500 points qui sont alors interpolés, soit directement dans l'espace



(a) Erreur de localisation dans S. Erreur moyenne : ILD = 2.86°(2.42°), ITD = 2.86°(5.30°).

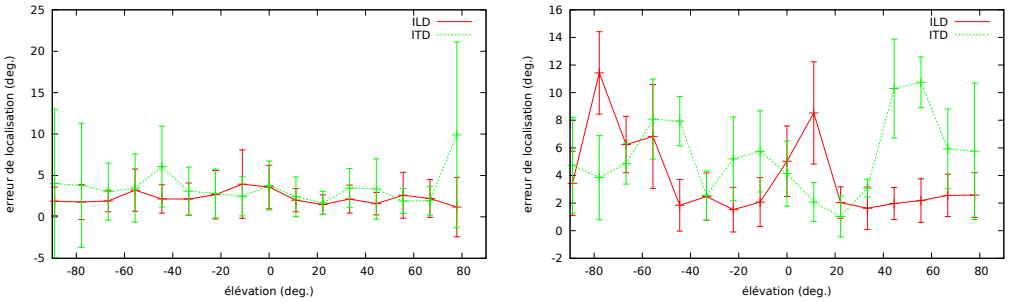
(b) Erreur de localisation dans R. Erreur moyenne : ILD = 2.91°(2.34°), ITD = 6.32°(6.83°).

FIGURE C.2 – Localisation en azimut par interpolation des k-ppv à partir des données CAMIL, à partir des vecteurs d'ILD (en rouge) et d'ITD (en vert). L'azimut est compris dans l'intervalle $[-180°, 180°]$ par pas de 2°, l'élévation est constante à 1°. Les barres d'erreurs représentent l'écart type. (a) Interpolation dans l'espace sensoriel S. (b) Interpolation dans l'espace de représentation R estimé par la méthode des LE. L'erreur moyenne obtenue et son écart-type sont précisés en légende pour chaque courbe.

sensoriel S, soit après projection dans l'espace de représentation R.

Concernant la localisation en azimut tout d'abord, dont les résultats sont détaillés Fig. C.2, nous observons globalement de bonnes performances. Dans l'espace S, l'erreur d'interpolation moyenne est identique en ILD et en ITD, atteignant 2.86° en moyenne. L'écart-type associé est cependant nettement plus important en ITD. Constatons de plus que la qualité de l'estimation est relativement homogène sur l'ensemble de l'espace azimutal et que nous n'observons pas ici l'effet de flou de localisation tel que rapporté chez l'homme (voir Fig. 2.2). Dans l'espace R cette fois, la localisation à base d'ILD est encore très bonne, avec une erreur moyenne de 2.91°, néanmoins les perfomances associées à l'ITD chutent (la moyenne passant de 2.86° à 6.32°) à cause d'erreurs d'estimation importantes pour des directions spécifiques, notemment autour des 90° où l'erreur atteint les 25°. Cette large erreur ne peut s'expliquer par un problème au niveau de l'indice d'ITD puisque l'estimation dans S s'effectue correctement, ni par un problème lié à la variété qui serait mal apprise. Ainsi nous en concluons - sur la base de ces résultats mais également d'expériences préliminaires non décrites ici - que cette erreur est liée à la méthode de projection P_e pour laquelle nous avons constaté une chute de performance lorsque la dimension de l'espace d'entrée est importante - 690 en ITD contre 30 en ILD, rappelons le.

Concernant maintenant la localisation en élévation, les résultats détaillés étant représentés Fig. C.3, nous observons là encore de bonnes performances d'autant plus que, comme nous l'avons dit, ce phénomène de localisation en élévation à partir d'indices binauraux n'est pas - ou très peu - rapporté dans la littérature, le problème de la localisation en élévation étant bien souvent traité sur la base de l'analyse explicite des HRTF. Ainsi, à partir d'une interpolation dans S, l'erreur moyenne atteint 2.29° en ILD et 3.60° en ITD et nous observons là encore une augmentation de l'écart-type associé entre ILD et ITD. Le passage de S à R entraîne quand à lui une augmentation de l'erreur d'estimation pour l'ILD plus marquée que dans la cas azimutal, mais moins marquée pour l'ITD. Comme nous l'avons expliqué précédemment, ces



(a) Erreur de localisation dans S. Erreur moyenne : ILD = 2.29°(2.49°), ITD = 3.60°(5.33°).

(b) Erreur de localisation dans R. Erreur moyenne : ILD = 3.93°(3.59°), ITD = 5.42°(3.82°).

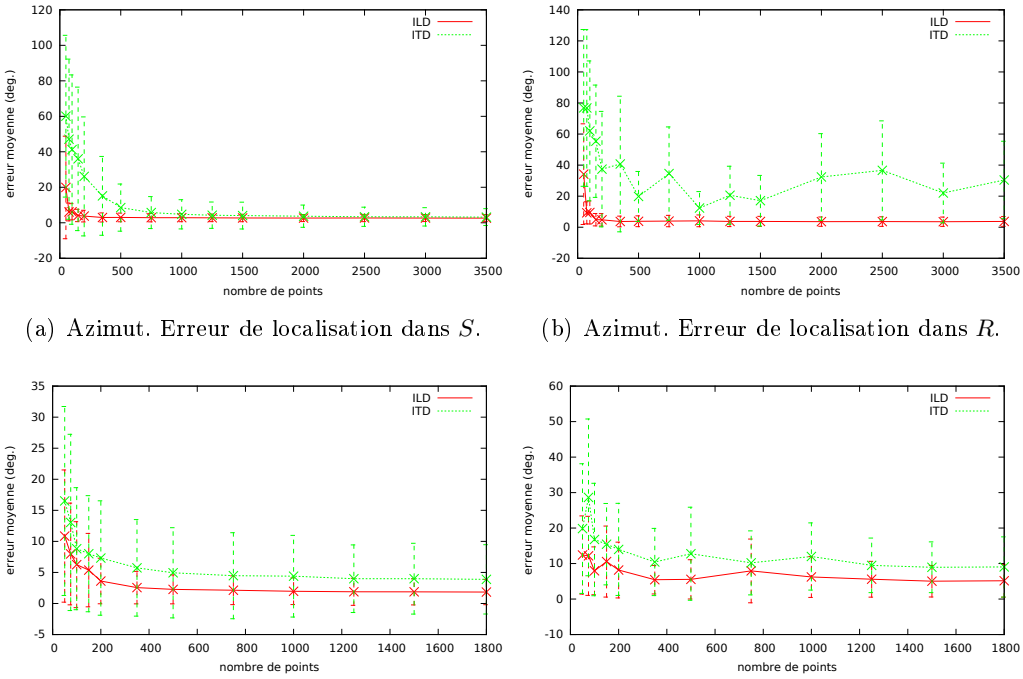
Figure C.3 – Localisation en élévation par interpolation des k-ppv à partir des données CAMIL, à partir des vecteurs d'ILD (en rouge) et d'ITD (en vert). L'élévation est comprise dans l'intervalle $[-90°, 90°]$ par pas de 2°, l'azimut est constant à 1°. Les barres d'erreurs représentent l'écart type. (a) Interpolation dans l'espace sensoriel S. (b) Interpolation dans l'espace de représentation R estimé par la méthode des LE. L'erreur moyenne obtenue et son écart-type sont précisés en légende pour chaque courbe.

bonnes performances de localisation en élévation s'expliquent à la fois par la densité de l'échantillonnage et par la présence implicite de non-linéarités liées aux HRTF et aux conditions acoustiques dans les indices d'ILD et d'ITD. Dans un expérience de localisation audio-visuelle, Youssef *et al.* (2012a) ont obtenu de bonnes performances de localisation en élévation, où la position de la source est exprimée dans l'espace image associé à une caméra. A partir d'un modèle d'apprentissage suprevisé à base de MLP, ils utilisent pour cela un modèle binaural à base de gammatone et des indices auditifs composés de la concaténation des énergies gauche et droite, donnant selon les auteurs de meilleurs résultats que l'ILD ou l'ITD.

C.2.3 Influence du nombre de points

Dans l'expérience précédente nous avons fixé arbitrairement le nombre d'échantillons à 1500, à la fois pour l'ensemble d'apprentissage et celui de test. Nous nous proposons de revenir ici plus en détail sur l'évolution de l'erreur moyenne de localisation en fonction du nombre d'échantillons composants l'ensemble d'apprentissage. La même méthodologie que précédemment est conservée, le nombre d'échantillons présents dans l'ensemble de test étant cette fois-ci fixé à 1000, la taille de l'ensemble d'apprentissage variant de 50 à 3600 pour l'azimut et à 1800 pour l'élévation, soit le nombre maximal de points générés à partir de la base CAMIL.

La Fig. C.4 présente le détail des résultats obtenus. Nous constatons dans tous les cas étudiés que les performances liées à l'ILD convergent plus rapidement que celles liées à l'ITD, de plus vers une erreur plus faible. Comme constaté précédemment, la localisation est moins précise dans R que dans S, et particulirement pour l'ITD. D'une manière générale, les 1500 échantillons que nous utilisions précédemment ne sont pas nécessaires, un ordre de grandeur de 500 points nous donnant des résultats proches des performances optimales. Enfin nous ne constatons pas de différence significative entre la vitesse de convergence de l'erreur entre R et S, les courbes suivant



(a) Azimut. Erreur de localisation dans S.

(b) Azimut. Erreur de localisation dans R.

(c) Elévation. Erreur de localisation dans S.

(d) Elévation. Erreur de localisation dans R.

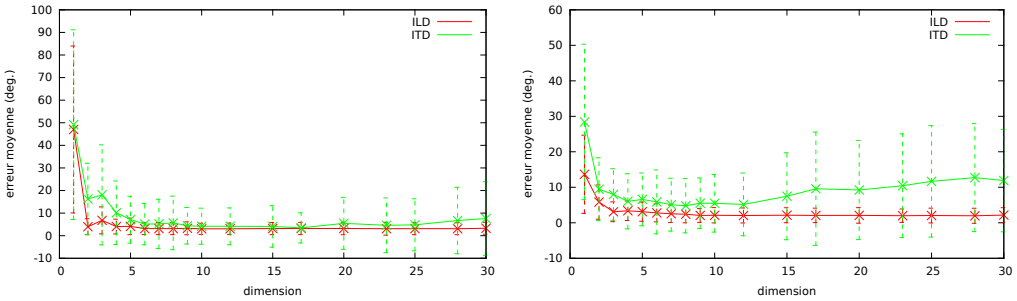
FIGURE C.4 – Erreur de localisation en fonction du nombre d'échantillons dans l'ensemble d'apprentissage, avec l'écart type associé, calculée à partir des indices d'ILD (rouge) et d'ITD (vert). (a et b) erreur moyenne en azimut. (c et d) erreur moyenne en élévation. (a et c) localisation dans S. (b et d) localisation dans R.

dans ces 2 cas des tendances similaires. Ce résultat est à noter car on pourrait *a priori* s'attendre à ce que l'apprentissage nécessite un échantillonnage moins dense dans R que dans S pour obtenir des performances comparables, du fait de la différence de dimensionnalité.

C.2.4 Influence de la dimension d'apprentissage

Nous avons constaté précédemment que les performances de localisation chutent quand l'estimation est effectuée dans R dont la dimension d_r était fixée à $d_r = 2$, respectivement aux performances obtenues dans S. Cette dernière expérience vise à évaluer l'influence de d_r sur les performances de localisation. Suivant les même conditions expérimentales que précédemment, nous considérons ici des ensemble d'apprentissage et de test tous deux composés de 1000 échantillons choisis d'une manière aléatoire et uniforme sur l'ensemble des vecteurs générée. Nous considérons là encore à la fois les performances en ILD et en ITD séparemment, de même pour l'azimut et l'élévation. La dimension d_r varie entre 1 et 30, soit la dimension maximale atteignable en ILD.

La Fig. C.5 présente le détail des résultats de cette expérience, qui semblent contre-intuitifs à plusieurs points de vue. Considérant tout d'abord la localisation en azimut, nous voyons que si l'ILD atteint comme on s'y attend des performances optimales dès $d_r = 2$ pour rester stable par la suite (avec néanmoins un léger rebond pour $d_r = 3$), cela n'est pas vérifié pour l'ITD dont l'erreur moyenne décroit jusqu'à $d_r = 6$ puis augmente à nouveau pour $d_r > 25$. Ce phénomène observé sur l'ITD en azimut est encore plus marqué lorsque l'on considère l'élévation. Ainsi son erreur



(a) Azimut. (b) Elévation.

Figure C.5 – Erreur de localisation dans R en fonction de sa dimension d_r, avec l'écart type associé, calculée à partir des indices d'ILD (rouge) et d'ITD (vert). (a) erreur moyenne en azimut. (b) erreur moyenne en élévation.

moyenne devient minimale vers $d_r = 6$ mais s'accroit de nouveau lorsque $d_r > 12$. Concerant pour finir le cas de l'ILD en élévation nous voyons que $d_r = 3$ est la dimension à partir de laquelle les performances deviennent stables, alors que la dimension théorique de l'espace auditif est ici égale à 1.

Cette expérience faisant varier la dimension d'apprentissage nous a donc montré des résultats différents de ceux attendus *a priori*. Comment expliquer cet écart ? Plusieurs causes peuvent être avancées. La forte courbure des variétés liées à l'élévation par exemple, peut être la cause d'un mauvais fonctionnement de la méthode de projection P_e ou de l'algorithme LE. Laflaquière (2013) a en effet reporté un disfonctionnement de sa méthode d'estimation de la dimension intrinsèque causé précisemment par une forte courbure locale des variétés étudiées. Quand aux performances obtenues en ITD, meilleures avec $d_k = 6$ que $d_k = 2$, cela peut éventuellement s'expliquer par la présence d'informations spectrales très riches, du fait de la grande dimension de l'espace d'entrée et de le sensibilité de l'ITD à ces informations, ce qui pourrait être investigué par une analyse plus en détail des résultats et par une estimation de la dimension intrinsèque des variétés par exemple.

Annexe D

Détails d'implémentation

Sommaire

Cette annexe présente quelques détails relatifs à l'implémentation du modèle binaural présenté au chapitre 4. Tout d'abord la bibliothèque C++ implémantant ce modèle est présentée dans sa globalité. Ensuite le détail de l'implémentation du modèle de Jeffress est proposé (voir pargraphe 4.2.1). Enfin nous présentons une implémentation électronique embarquée du filtrage gammatone (voir paragraphe 4.1.2).

D.1 Implémentation des traitements binauraux

Ce paragraphe présente l'implémentation des traitements binauraux introduis au chapitre 4 sous la forme d'une bibliothèque C++. L'intégralité des résultats présentés dans cette thèse sont issus de cette implémentation. Nous présentons tout d'abord le contexte applicatif puis l'architecture générale de cette bibliothèque, enfin nous présentons des mesures de temps d'éxécution démontrant son utilisabilité en temps réel.

D.1.1 Contexte applicatif

L'objectif de cette bibliothèque est de fournir des traitements auditifs efficaces au travers un accès simple à leur configuration et leur utilisation. Bien que l'intérêt premier de cette bibliothèque soit la robotique, les différents traitements peuvent être accédés selon 3 scénarios applicatifs présentés Fig. D.1. Les traitements binauraux

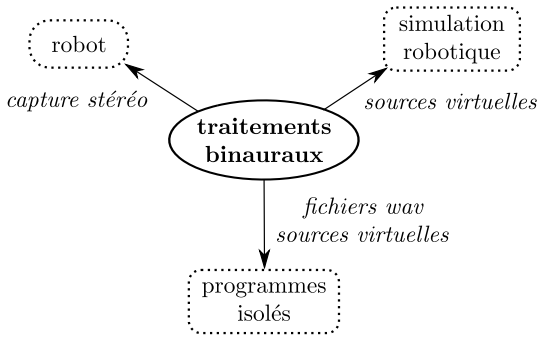


FIGURE D.1 – Contexte applicatif de la bibiliothèque, en conditions actives en robotique ou en simulation, en conditions passives dans des programmes isolés.

sont en effet décorrélés de l'application finale, un même modèle pouvant être utilisé de manière transparente dans ces 3 contextes :

— Robotique : le signal d'entrée est capturé en temps réel depuis les microphones du robot, les sorties sont des commandes motrices. Utilisé au paragraphe 6.
— Simulation robotique : robot et sources sonores sont simulés et placés dans un environnement virtuel. Utilisé au paragraphe 6.1.1.
— Programmes isolés : analyse de signaux binauraux virtuels ou préalablement enregistrés. Utilisé aux paragraphes 4.3 et 5.3 notamment.

Notons pour finir que ce modèle binaural peut également servir de base à d'autres travaux, au-delà de la « simple » localisation de sources sonores. Nous avons ainsi utilisé ce modèle dans un contexte de perception visuo-auditive en collaboration avec N'Guyen (2010). Ce dernier expose les détails de l'expérience consistant à la fusion des modalités auditives et visuelles dans un modèle de SC. L'implémentation des filtres gammatone que nous proposons est également utilisée dans des expériences de fusion multimodale bioinspirée (Pitti *et al.* , 2012a,b). Ces deux exemples nous permettent ainsi d'illustrer l'intérêt de disposer d'un système auditif artificiel adapté à la robotique et utilisé en prétraitement pour des tâches plus complexes.

D.1.2 Architecture

La bibliothèque se décompose en trois sous-ensembles que sont (1) les différents traitements proposés et qui constituent le coeur de la bibliothèque, (2) des modules d'entrée et (3) des modules de sortie. Nous revenons sur chacun de ces éléments ci-dessous.

Traitements Les traitements de base proposés sont :

— Les modèles cochléaires gammatone et de Lyon,
— La représentation impulsionnelle,
— L'intégration temporelle (comprend le sous-échantillonnage),
— L'extraction des fronts d'ondes,
— L'extraction des indices d'ILD et d'ITD.

Du point de vue de la conception logicielle, chaque classe de traitement hérite de la classe `ProcessingElement`. Cette classe abstraite, détaillée par la Fig. D.2, fourni les opérations de base accessibles à tous les blocs de traitements. En plus de définir la fréquence d'échantillonage de travail ainsi que le nombre de canaux, cette classe offre les fonctions `compute()` permettant de cacluler la sortie du filtre en fonction de

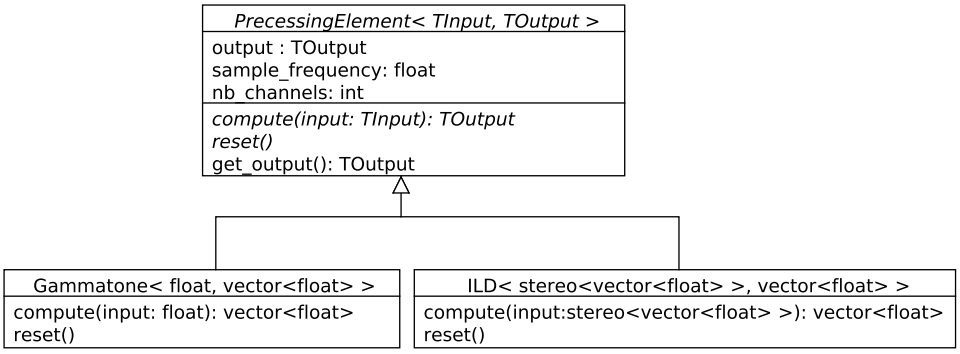


Figure D.2 – Diagramme UML simplifié de la classe abstraite `ProcessingElement` et exemple d'héritage vers les classes `Gammatone` et `ILD`. `ProcessingElement` est paramétrée par les types génériques `TInput` et `TOuput` correspondant respectivement aux types des données en entrée et en sortie. Les classes filles doivent spécifier ces types et implémenter les fonctions abstraites `compute(input: TInput): TOutput` et `reset()`. La structure générique `stereo`, composée d'un champ `left` et d'un champ `right`, permet de représenter une donnée binaurale. Les attributs et fonctions spécifiques aux classes filles ne sont pas spécifiés.

l'entrée et de son état interne, et la fonction `reset()` permettant de mettre à zéro l'ensemble des attributs de données d'une instance de cette classe. De plus, la classe `ProcessingElement` est responsable de l'allocation et de la gestion de son propre état de sortie, la transmission des données d'un traitement à l'autre s'effectuant donc par référence constante et non par copie. Considérant ainsi deux instances `a` et `b` de `ProcessingElement` composant 2 traitements effectués à la chaîne, la sortie `out_b` de `b` s'exprime en fonction de l'entrée `in_a` de `a` comme `out_b = b.compute( a.compute( in_a ) )`.

Entrées Les différents paramètres du modèle sont chargés depuis un fichier externe au travers une classe de configuration. Les signaux sonores accessibles au système sont quand à eux accessibles par 3 entrées différentes (voir Fig. D.1) :

— L'enregistrement à partir de microphones et la capture en temps réel,
— Le chargement d'un fichier depuis un fichier externe,
— La génération de sources virtuelles (bruit blanc, sinusoïde, etc).

Dans un contexte de simulation, chaque source simulée, qu'elle soit générée ou provienne d'un fichier mono, est ensuite spatialisée par atténuation en distance et simulation de HRTF (voir paragraphe 4.1.1, modèles 3 et 4). Le simulateur peut prendre en compte un nombre arbitraire de source mais ni l'effet Doppler, associé aux mouvements relatifs des sources, ni l'acoustique de la pièce ne sont modélisées, limitant ces simulations au contexte anéchoïque.

Sorties Les différentes sorties accessibles au système sont :

— Des commandes motrices asservissant le cou et le déplacement d'un robot, que ce soit sur plateforme réelle on en simulation,
— L'écriture de fichiers de *log* permettant une analyse ultérieure des données brutes,
— La visualisation graphique en temps réel de l'état du système (signaux temporels, cochléogrammes, indices d'ILD et d'ITD, proprioception).

f_s (kHz)	t_1 (s)	t_2 (s)
20	3.49	1.79
44.1	7.71	3.9

TABLE D.1 – Temps d'exécution du modèle binaural pour 5 s de signal échantillonné à $f_s = 20$ ou 44.1 kHz. t_1 est le temps obtenu à partir du modèle de Jeffress non intégré, t_2 à partir du modèle intégré en fréquence. La proportion de temps occupée par l'estimation de l'ITD est de 70.7% pour t_1 contre 46.3% pour t_2.

D.1.3 Performance temps-réel

Nous considérons les temps d'exécution du modèle binaural associé à 2 versions différentes du modèle de Jeffress : la version « originale » telle que proposée au paragraphe 4.2.1 et une version proposant une intégration en fréquence de l'ITD. En effet la dimension de sortie du modèle de Jeffress peut être élevée (690 pour les expériences du paragraphe 5.3) et son intégration temporelle devenir coûteuse. Or si l'objectif est uniquement l'estimation de l'ITD moyen sur l'ensemble du spectre, comme au paragraphe 4.3 par exemple, il est préférable d'effectuer l'intégration temporelle après l'intégration fréquentielle. Ceci réduit en effet la complexité de l'intégration temporelle d'un facteur égal au nombre de canaux fréquentiels utilisés.

Les temps d'exécution relevés dans la Table D.1 sont obtenus sur un processeur datant de 2009. L'évaluation est effectuée sur un modèle complet incluant deux systèmes périphériques (cochlées à 30 canaux et génération d'impulsions), un module d'ILD et un module d'ITD (avec extraction des fronts d'onde, $d_{inter} = 0.19$ m). Les mesures prennent en compte uniquement les traitements auditifs, le temps d'exécution associé aux entrées et sorties n'étant pas considéré. L'entrée est un signal stereo aléatoire d'une durée de 5 s échantillonné à $f_s = 20$ ou 44.1 kHz pour les 2 versions du modèle de Jeffress. Parmi les 4 configurations mesurées, seule la plus complexe (44.1 kHz, version non-intégrée) n'est pas appliquable en temps réel sur notre plateforme puisqu'elle met 7.71 s à traiter un signal de 5 s, 70.7% du temps de calcul étant consacré à l'estimation de l'ITD.

D.2 Implémentation du modèle de Jeffress

Ce paragraphe détaille l'implémentation que nous proposons du modèle de Jeffress décrit au paragraphe 4.2.1. Nou présentons tout d'abord le principe soutenant cette implémentation puis nous introduisons le détail de l'algorithme et en proposons un pseudo-code.

D.2.1 Principe

Ce modèle, comme nous l'avons décrit au paragraphe 4.2.1, permet d'extraire l'ITD grâce à deux lignes de délais à l'intersection desquelles sont détectées des coïncidences. C'est l'accumulation temporelle de ces coïncidences pour les différents canaux fréquentiels qui donne l'estimation de l'ITD. Le modèle de Jeffress est utilisé dans notre système binaural en sortie du bloc d'extraction des fronts d'onde, l'algorithme reçoit donc en entrée les valeurs des trains d'impulsions gauche et droit pour chaque canal fréquentiel à un pas de temps donné. Il retourne en sortie une matrice délai-fréquence dans laquelle est stockée le nombre de coïncidences détectées pour

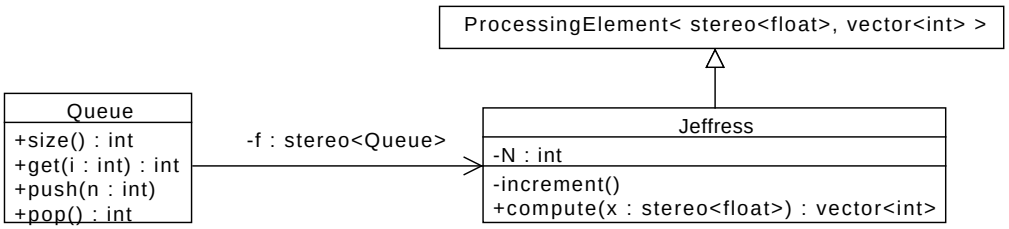


Figure D.3 – Diagramme UML simplifié du code implémentant le modèle de Jeffress pour un unique canal fréquentiel. La structure générique `stereo`, composée d'un champ `left` et d'un champ `right`, permet de représenter une donnée binaurale. Les attributs de `Queue` ne sont pas spécifiés, de même que les membres de `ProcessingElement`, dont hérite la classe `Jeffress` (voir annexe D.1). `N` correspond au nombre d'échantillons relatif au délai maximal (voir Eq. 4.12).

chaque délai et chaque canal fréquentiel durant le temps correspondant à la durée d'intégration (il s'agit donc d'une matrice d'entiers). Puisque les traitements effectués sont identiques pour tous les canaux fréquentiels, nous ne présentons ici qu'une version monocanal de l'algorithme, son extension au cas multicanal étant triviale. Un diagramme UML schématisant la structure de cette implémentation, composée des classes `Queue` et `Jeffress`, est présenté Fig. D.3.

La classe `Jeffress`, qui hérite de `ProcessingElement` (voir l'annexe D.1), se focalise exclusivement sur la composante temporelle de l'entrée. L'amplitude des impulsions n'est pas prise en compte. De plus l'algorithme ne modélise pas directement les lignes de délai mais les représente de manière indirecte, en stockant le délai associé à chaque impulsion transitant dans ces lignes. Deux files f_l et f_r sont donc utilisées à cet effet, associées aux impulsions provenant de gauche et de droite respectivement. Conformément à la notation de la Fig. D.3, nous avons donc $f = (f_l, f_r)$.

D.2.2 Algorithme

Ce paragraphe détaille successivement l'implémentation de la classe `Queue` et des fonctions `Jeffress::increment()` et `Jeffress::compute()`.

La classe `Queue` La classe `Queue` est une structure de donnée classique : la file d'entiers. En pratique le code de la classe `Queue` repose sur la classe `std::deque`, l'implémentation des files proposée par la bibliothèque standard de C++. Elle implémente les fonctions membres suivantes :

- `size()` renvoie le nombre d'éléments dans la file ;
- `operator[i]` accède en lecture ou écriture à l'élément d'index `i`$\in$ [0,`size()`-1] ;
- `push(n)` ajoute l'élément `n` en queue de file ;
- `pop()` supprime l'élément en tête de file et renvoie sa valeur.

La fonction `Jeffress::increment` A chaque itération, la fonction `Jeffress::increment` incrémente chacun des délais contenus dans les deux files et supprime les délais dont la valeur est supérieure au délai maximum `N`, comme le détaille l'Algo. 1. Ce retard maximal s'exprime en nombre d'échantillons – un entier donc – et se calcule en fonc-

Algorithme 1 Implémentation de la fonction `Jeffress::increment`.

pour $i = 0$ à $f_l.size() - 1$ **faire**
  $f_l[i] \leftarrow f_l[i] + 1$
fin pour
si $f_l[0] > N$ **alors**
  $f_l.pop()$
fin si

pour $i = 0$ à $f_r.size() - 1$ **faire**
  $f_r[i] \leftarrow f_r[i] + 1$
fin pour
si $f_r[0] > N$ **alors**
  $f_r.pop()$
fin si

tion de la fréquence d'échantillonnage, de la distance interaurale et de la vitesse du son, comme décrit au paragraphe 4.2.1.

La fonction `Jeffress::compute` Maintenant que la modélisation des lignes de délai est en place, il s'agit à présent de détecter des coïncidences entre ces lignes à gauche et à droite. C'est le rôle de la fonction `Jeffress::compute`, qui est appelée à chaque échantilon. Puisque nous nous plaçons ici dans un cas monocanal, l'entrée de cette fonction consite donc simplement en deux nombres réels x_l et x_r représentant, pour un pas de temps donné, la valeur des impulsions à gauche et à droite respectivement. Conformément à la Fig. D.3 nous avons donc $x = (x_l, x_r)$. Le pseudo-code de cette fonction est présenté par l'Algo. 2. Selon les valeurs de x, l'algorithme distingue les 4 cas suivants :

1. Impulsions simultanées à gauche et à droite ($x_l \neq 0$ et $x_r \neq 0$) : Si les deux files sont vides, *i.e.* si aucune impulsion ne se propage dans les lignes de délai, une coincidence est détectée pour un délai nul. Si au moins une des deux file n'est pas vide une coïncidence est détectée pour le délai correspondant à la tête de file, ainsi une seule des deux impulsions en entrée est « consomée », l'autre est rajoutée en queue de file avec un délai nul ;
2. Impulsion à gauche uniquement ($x_l \neq 0$ et $x_r = 0$) : Si la file de droite est vide, un délai nul est rajouté en queue de la file de gauche. Sinon une coincidence est détectée pour un délai correspondant à celui de la tête de la file droite, celui-ci étant ensuite supprimé de sa file ;
3. Impulsion à droite uniquement ($x_l = 0$ et $x_r \neq 0$) : De manière symmétrique, le même traitement que pour le cas précédent est effectué à partir de la file gauche ;
4. Pas d'impulsion ($x_l = 0$ et $x_r = 0$) : Ceci correpond au cas le plus simple et le plus fréquent. Aucun traitement n'est effectué, une entrée nulle implique une absence de coincidence en sortie.

Une fois que la détection de coïncidence a eu lieu, les délais présents dans les files sont incrémentée par la fonction `Jeffress::increment` et le vecteur s est renvoyé. Celui-ci contient au plus une coïncidence, c'est à que $\sum_i s_i$ est soit nul soit unitaire.

Le code proposé s'avère efficace puisqu'il tire directement profit de la structure impulsionnelle de l'entrée. Ainsi aucune multiplication n'est effectuée, l'agorithme

Algorithme 2 Implémentation de la fonction `Jeffress::compute`.

Entrée : $x = (x_l, x_r)$ avec $x \in \mathbb{R}^2$,
Sortie : $s \in \mathbb{N}^{2N+1}$ initialisé comme vecteur nul

Cas 1 : impulsions simultanées à gauche et à droite
si $x_l \neq 0$ **et** $x_r \neq 0$ **alors**
  si $f_l.size() = 0$ **et** $f_r.size() = 0$ **alors**
    $s(N) = 1$
  sinon
    si $f_r.size() \neq 0$ **alors**
      $s[N + f_r.pop()] \leftarrow 1$
      $f_r.push(0)$
    sinon
      $s[N - f_l.pop()] \leftarrow 1$
      $f_l.push(0)$
    fin si
  fin si
fin si

Cas 2 : impulsion à gauche
si $x_l \neq 0$ **et** $x_r = 0$ **alors**
  si $f_r.size() \neq 0$ **alors**
    $s[N + f_r.pop()] \leftarrow 1$
  sinon
    $f_r.push(0)$
  fin si
fin si

Cas 3 : impulsion à droite
si $x_l = 0$ **et** $x_r \neq 0$ **alors**
  si $f_l.size() \neq 0$ **alors**
    $s[N - f_l.pop()] \leftarrow 1$
  sinon
    $f_l.push(0)$
  fin si
fin si

$increment()$
renvoyer s

se composant exclusivement de mouvements en mémoire et d'incrémentations. A ce titre cet algorithme propose une alternative intéressante et tout à fait opposée aux méthodes basées sur la corrélation.

D.3 Implémentation embarquée du filtrage gammatone

Le modèle de cochlée à base de filtres gammatone présenté au pargraphe 4.1.2 a fait l'objet d'une implémentation numérique en *hardware*. Cette implémentation est

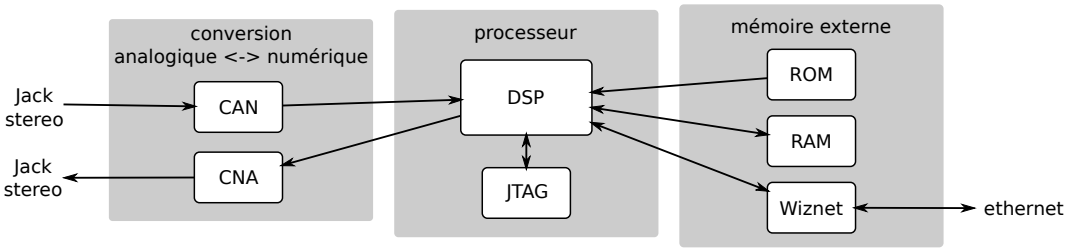


FIGURE D.4 – Schématique générale de la carte cochlée.

motivée par la nécessité, dans le domaine de la robotique, de disposer de solutions de traitements embarquées à basse consommation (voir à ce sujet le paragraphe 3.1.3). Ainsi une solution électronique présentée ci-dessous a été proposée par la société Brain Vision Systems, le travail effectué durant cette thèse consistant à l'écriture du pilote matériel de la carte et à une implémentation des filtres gammatone optimisée pour l'architecture spécifique du processeur de signal numérique embarqué sur la carte. Le signal de sortie est transmis au client via le réseau local par une connection ethernet. Cette solution hardware peut se décomposer en 3 sous-ensembles, comme illustré Fig. D.4, qui sont présentés successivement.

D.3.1 Acquisition

L'entrée audio de la carte est une entrée analogique matérialisée par un connecteur Jack stereo 3.5 mm relié à un convertisseur analogique-numérique *Analog Device AD1937*. L'échantillonnage du signal est effectué à 48 kHz pour une résolution de 24 bits. Une fois numérisé, le signal audio est transmis au processeur via son interface d'entrée. Un Jack de sortie est également présent sur la carte. Celui-ci permet soit de récupérer le signal d'entrée éventuellement filtré, soit d'accéder à la sortie d'un canal fréquentiel du banc de filtres.

D.3.2 Processeur

Le processeur de signal numérique *Analog Device AD21369* constitue le coeur de la carte. Sa fréquence de travail est fixée à 324 MHz. Le programme exécuté par ce processeur permet d'estimer les cochléogrammes gauche et droite en temps réel. Dans le but de minimiser la taille des trames en sorties, la représentation impulsionnelle est également implémentée (voir le paragraphe 4.1.3). En effet, puisqu'une impulsion est suivie d'au moins une valeur nulle (beaucoup plus en pratique), la sortie est compressée de la manière suivante, indépendemment sur chacun des canaux à gauche ou à droite : chaque impusion détectée est représentée par le couple (n, s) où n est le nombre de valeurs nulles précédant l'impulsion détectée et s est sa valeur. Une fois que la trame a atteint une taille critique celle-ci est transmise au client via le contrôleur ethernet (voir ci-dessous).

La carte est également équipée d'une sonde JTAG, qui est un outil de dévelloppement et de *debug*. Cela permet en effet de charger un programme compilé dans la mémoire interne du processeur, ou dans la ROM associée. Cette sonde permet également d'accéder à l'état interne du processeur durant l'exécution d'un programme.

D.3.3 Mémoire externe

La mémoire externe se décompose en 3 entités : la ROM stockant le programme compilé, la RAM servant au stockage des données temporaires et un contrôleur ethernet *Wiznet W5100*. Ce contôleur permet de transmettre les trames de sorties compressées au client au travers une interface ethernet et le réseau local. Enfin l'opération de base effectuée par le client (en pratique un ordinateur de bureau) consiste en la récupération des trames de sorties et en leur décompression et resynchronisation.

Bibliographie

Alameda-Pineda, X, Khalidov, V., Horaud, R., & Forbes, F. 2011. Finding audio-visual events in informal social gatherings. *Proceedings of the 13th international conference on multimodal interfaces*, 247–254.

Algazi, V.R., Duda, R.O., Morrison, R.P., & Thompson, D.M. 2001a. Structural composition and decomposition of HRTFs. *Ieee workshop on applications of signal processing to audio and acoustics*, 103–106.

Algazi, V.R., Duda, R.O., Thompson, D.M., & Avendano, C. 2001b. The CIPIC HRTF database. *Ieee workshop on applications of signal processing to audio and acoustics*, 99–102.

Algazi, V.R., Duda, R.O., Duraiswami, R., Gumerov, N.A., & Tang, Z. 2002. Approximating the head-related transfer function using simple geometric models of the head and torso. *Journal of the acoustical society of america*, **112**(5), 2053–2064.

Allen, J.B., & Berkley, D.A. 1979. Image method for efficiently simulating small-room acoustics. *Journal of the acoustical society of america*, **65**(4), 943–950.

Andermann, M.L., & Moore, C.I. 2008. Mechanical resonance enhances the sensitivity of the vibrissa sensory system to near-threshold stimuli. *Brain research*, **1235**(Oct.), 74–81.

Arabzadeh, E., Panzeri, S., & Diamond, M.E. 2004. Whisker vibration information carried by rat barrel cortex neurons. *Journal of neuroscience*, **24**(26), 6011–20.

Arabzadeh, E., Zorzin, E., & Diamond, M.E. 2005. Neuronal encoding of texture in the whisker sensory pathway. *Plos biology*, **3**(1), e17.

Argentieri, S., & Danès, P. 2007. Broadband variations of the MUSIC high-resolution method for sound source localization in robotics. *Ieee international conference on intelligent robots and systems*, Oct., 2009–2014.

Argentieri, S., Portello, A., Bernard, M., Danès, P., & Gas, B. 2013. Binaural systems in robotics. *Chap. 9 of :* Blauert, J. (ed), *The technology of binaural listening.* Springer.

Ashmead, D.H., Davis, D.L., & Northington, A. 1995. Contribution of listeners approaching motion to auditory distance perception. *Journal of experimental psychology : Human perception and performance*, **21**(2), 239–56.

Auvray, M., & O'Regan, J.K. 2003. L'influence des facteurs sémantiques sur la cécité aux changements progressifs dans les scènes visuelles. *L'année psychologique*, **103**, 9–32.

Auvray, M., Hanneton, S., Lenay, C., & O'Regan, J.K. 2005. There is something out there : distal attribution in sensory substitution, twenty years later. *Journal of integrative neuroscience*, **4**(4), 505–21.

Aytekin, M., Moss, C.F., & Simon, J.Z. 2008. A sensorimotor approach to sound localization. *Neural computation*, **20**(3), 603–635.

Bach-y Rita, P., & Kercel, S.W. 2003. Sensory substitution and the human–machine interface. *Trends in cognitive sciences*, **7**(12), 541–546.

Bach-y Rita, P., Collins, C., Saunders, F.A., White, B., & Scadden, L. 1969. Vision substitution by tactile image projection. *Nature*, **221**(5184), 963–964.

Backman, J., & Karjalainen, M. 1993. Modelling of human directional and spatial hearing using neural networks. *Ieee international conference on acoustics, speech, and signal processing*, 125–128.

Baghani, A., & Araabi, B.N. 2006. Improved Laplacian Eigenmaps. *International csi computer conference.*

Baker, C. 1977. *The numerical treatment of integral equations.* Clarendon Press, Oxford.

Batteau, D W. 1967. The role of the pinna in human localization. *Proceedings of the royal society of london*, **168**(11), 158–80.

Beitel, R E. 1999. Acoustic pursuit of invisible moving targets by cats. *Journal of the acoustical society of america*, **105**(6), 3449–53.

Belin, P, Zatorre, RJ, Lafaille, P, Ahad, P, & Pike, B. 2000. Voice-selective areas in human auditory cortex. *Nature*, **403**(6767), 309–12.

Belkin, M., & Niyogi, P. 2002. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems*, **1**, 585–592.

Belkin, M., & Niyogi, P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, **15**(6), 1373–1396.

Bengio, Y., Paiement, J.F., Vincent, P., Delalleau, O., Le Roux, N., & Ouimet, M. 2003. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. *Advances in neural information processing systems*, **16**, 177–184.

Bergan, J., & Knudsen, E. 2008. Auditory map plasticity in juvenile and adult owls. *In :* Dallos, P., & Oertel, D. (eds), *The senses : a comprehensive reference.* Elsevier.

Berglund, E., & Sitte, J. 2005. Sound source localisation through active audition. *Ieee international conference on intelligent robots and systems*, 653–658.

Berglund, E., Sitte, J., & Wyeth, G. 2008. Active audition using the parameter-less self-organising map. *Autonomous robots*, **24**(4), 401–417.

Bernard, M., N'Guyen, S., Pirim, P., Gas, B., & Meyer, J.A. 2010a. Phonotaxis behavior in the artificial rat Psikharpax. *International symposium on robotics and intelligent sensors*, 118–122.

Bernard, M., Pirim, P., de Cheveigné, A., & Gas, B. 2012. Sensorimotor learning of sound localization from an auditory evoked behavior. *Ieee international conference on robotics and automation*, 91–96.

Bernard, Mathieu, Pirim, Patrick, Meyer, J.A., Gas, Bruno, N'Guyen, S., & Guillot, A. 2010b. A supramodal vibrissa tactile and auditory model for texture recognition. *International conference on simulation of adaptive behavior*, 188–198.

Beyer, R.P. 1992. A computational model of the cochlea using the immersed boundary method. *Journal of computational physics*, **162**, 145–162.

Bidet-Caulet, A., & Bertrand, O. 2009. Neurophysiological mechanisms involved in auditory perceptual organization. *Frontiers in neuroscience*, **3**(2), 182.

Blauert, J. 1997. *Spatial hearing.* MIT Press.

Blauert, J. 2013. *The technology of binaural listening.* Springer.

Braitenberg, V. 1986. *Vehicles : Experiments in synthetic psychology.* MIT Press.

Bregman, A.S. 1999. *Auditory scene analysis.* MIT Press.

Breteau, B., Argentieri, S., Zarader, J.-L., Zefeng, Wang, & Youssef, K. 2010. Binaural speaker recognition for humanoid robots. *Ieee international conference on robotics and biomimetics*, 1405–1410.

Bronkhorst, Adelbert W. A.W. 2000. The Cocktail Party Phenomenon : A Review of Research on Speech Intelligibility in Multiple-Talker Conditions. *Acustica*, **86**(1), 117–128.

Brown, G.J., Ferry, R.T., & Meddis, R. 2010. A computer model of auditory efferent suppression : implications for the recognition of speech in noise. *Journal of the acoustical society of america*, **127**(2), 943–54.

Brucke, M., Nebel, W., Schwarz, A., Mertsching, B., Hansen, M., & Kollmeier, B. 1998. Digital VLSI-implementation of a psychoacoustically and physiologically motivated speech preprocessor. *Proceedings of the nato advanced study institute on computational hearing*, 157–162.

Brucke, M, Schulz, A, & Nebel, W. 1999. Auditory signal processing in hardware : A linear gammatone filterbank design for a model of the auditory system. *International workshop in field-programmable logic and applications.*

Brutti, A., Omologo, M., & Svaizer, P. 2008. Comparison between different sound source localization techniques based on a real data collection. *International conference on hands-free speech communication and microphone arrays*, 69–72.

Calmes, L. 2009. *Biologically inspired binaural sound source localization and tracking for mobile robots.* Ph.D. thesis.

Calmes, L., Lakemeyer, G., & Wagner, H. 2007. Azimuthal sound localization using coincidence of timing across frequency on a robotic platform. *Journal of the acoustical society of america*, **121**(4), 2034–2048.

Caluwaerts, K, Staffa, M, N'Guyen, S., Grand, C, Dollé, L, Favre-Félix, a, Girard, B, & Khamassi, M. 2012. A biologically inspired meta-control navigation system for the Psikharpax rat robot. *Bioinspiration & biomimetics*, **7**(2), 025009.

Cariani, P. 2011. Jeffress model. *Scholarpedia*, **6**(7), 2920.

Casseday, J.H., Fremouw, T., & Covey, E. 2001. The inferior colliculus : a hub for the central auditory system. *In :* Oertel, D., Fay, R.R., & Popper, A.N. (eds), *Integrative functions in the mammalian auditory pathway.* Springer Handbook of Auditory Research.

Chan, V., Liu, S.C., & van Schaik, A. 2007. AER EAR : A matched silicon cochlea pair with address event representation interface. *Ieee transactions on circuits and systems*, **54**(1), 48–59.

Chaumette, F., & Hutchinson, S. 2006. Visual servo control. I. Basic approaches. *Robotics & automation magazine, ieee*, 82–90.

Chen, Z. 2003. *An odyssey of the cocktail party problem.* Tech. rept. Adaptive Systems Lab, McMaster University.

Cheng, C.I., & Wakefield, G.H. 2001. Moving sound source synthesis for binaural electroacoustic eusic using interpolated head-related transfer functions. *Computer music journal*, **25**(4), 57–80.

Cherry, E.C. 1953. Some experiments on the recognition of speech with one and with two ears. *Journal of acoustical society of america*, **25**(5), 975–980.

Cooke, M. 1993. *Modelling auditory processing and organisation.* Cambridge University Press.

Cooke, M, Lu, Y.-C., & Horaud, R. 2007. Active hearing, active speaking. *International symposium on auditory and audiological research*, 1–12.

Corey, C., & Wakefield, G.H. 2001. Introduction to head-related transfer functions (HRTFs) : representations of HRTFs in time, frequency, and space. *Journal of the audio engineering society*, **49**(4), 231–249.

Couverture, C., & Gas, B. 2009. Extracting space dimension information from the auditory modality sensori-motor flow using a bio-inspired model of the cochlea. *Ieee international conference on intelligent robots and systems*, 2742–2747.

Damaske, P., & Wagener, B. 1969. Investigations of directional hearing using a dummy head. *Acustica*, **22**, 30–35.

Damper, R, & French, R. 2003. Evolving spiking neuron controllers for phototaxis and phonotaxis. *Applications of evolutionary computing - lecture notes in computer science*, **2611**, 616–625.

de Cheveigné, A. 2004. Structure du système auditif.

Deleforge, A., & Horaud, R. 2011. *Learning the direction of a sound source using head motions and spectral features.* Tech. rept. February. INRIA.

Deleforge, A., & Horaud, R. 2012. A latently constrained mixture model for audio source separation and localization. *Acm/ieee international conference on human robot interaction*, **7191**, 431–438.

Devore, S., Ihlefeld, A., Shinn-Cunningham, B.G., & Delgutte, B. 2007. Neural and behavioral sensitivities to azimuth degrade with distance in reverberant environments. *Chap. 7, pages 505–516 of :* Kollmeier, B., Klump, G., Hohmann, V., Langemann, U., Mauermann, M., Uppenkamp, S., & Verhey, J. (eds), *Hearing–from sensory processing to perception.* Springer.

Doya, K. 1999. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex ? *Neural networks*, **12**(7-8), 961–974.

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A.J., & Shamma, S.A. 2009. Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron*, **61**(2), 317–329.

Faller, C., & Merimaa, J. 2004. Source localization in complex listening situations : Selection of binaural cues based on interaural coherence. *Journal of the acoustical society of america*, **116**(5), 3075–3089.

Fay, R.R., & Popper, A.N. 2010. *Computational models of the auditory system.* Springer Handbook of Auditory Research.

Fend, M., Bovet, S., Yokoi, H., & Pfeifer, R. 2003. An active artificial whisker array for texture discrimination. *Ieee international conference on intelligent robots and systems*, **2**, 1044–1049.

Fend, Miriam. 2005. Whisker-based texture discrimination on a mobile robot. *Advances in artificial life - lecture notes in computer science*, **3630**, 302–311.

Ferry, R.T., & Meddis, R. 2007. A computer model of medial efferent suppression in the mammalian auditory system. *Journal of the acoustical society of america*, **122**(6), 3519–26.

Fettiplace, R., & Fuchs, P. 1999. Mechanisms of hair cell tuning. *Annual review of physiology*, **61**(1), 809–34.

Fettiplace, R, & Hackney, C.M. 2006. The sensory and motor roles of auditory hair cells. *Nature reviews neuroscience*, **7**, 19–29.

Finger, H., Liu, S.C., Ruvolo, P., & Movellan, J.R. 2010. Approaches and databases for online calibration of binaural sound localization for robotic heads. *Ieee/rsj international conference on intelligent robots and systems*, 4340–4345.

Fitzpatrick, D. 2002. Transformations in processing interaural time differences between the superior olivary complex and inferior colliculus : Beyond the Jeffress model. *Hearing research*, **168**(1-2), 79–89.

Fix, E., & Hodges, J.L. 1951. *Nonparametric discrimination : consistency properties.* Tech. rept. May. USAF School of Aviation Medicine, Randolph Field, Texas.

Frolov, A.A. 2011. Physiological basis of 3-D external space perception : approach of Henri Poincaré. *In : History of the neurosciences in france and russia.* Hermann.

Garcia, B. 2013. *Compte rendu de stage : Etude d'un algorithme de perception active pour l'audition binaurale.* Tech. rept. Institut des systèmes intelligents et de robotique.

Gardner, W. 1994. *HRTF measurements of a KEMAR dummy-head microphone.* Tech. rept. 6. Media Laboratory, Massachusetts Institute of Technology.

Ghitza, O. 1994. Auditory models and human performance in tasks related to speech coding and speech recognition. *Ieee transactions on speech and audio processing*, **2**(1), 115–132.

Givelberg, E, & Bunn, J. 2003. A comprehensive three-dimensional model of the cochlea. *Journal of computational physics*, **191**, 377–391.

Givelberg, E, Bunn, J, & Rajan, M. 2001. *Detailed simulation of the cochlea : Recent progress using large shared memory parallel computers.* Tech. rept. California Institute of Technology.

Glackin, B., Wall, J.A., McGinnity, T.M., Maguire, L.P., & McDaid, L.J. 2010. A spiking neural network model of the medial superior olive using spike timing dependent plasticity for sound localization. *Frontiers in computational neuroscience*, **4**(Jan.), 1–16.

Glasberg, B., & Moore, B. 1990. Derivation of auditory filter shapes from notched-noise data. *Hearing research*, **47**, 103–138.

Goodman, D.F.M., & Brette, R. 2010a. Learning to localise sounds with spiking neural networks. *Advances in neural information processing systems*, 1–9.

Goodman, D.F.M., & Brette, R. 2010b. Spike-timing-based computation in sound localization. *Plos computational biology*, **6**(11), e1000993.

Grubbs, F. E. 1969. Procedures for detecting outlying observations in samples. *Technometrics*, **11**, 1–21.

Guillot, A, & Meyer, J.A. 2001. The animat contribution to cognitive systems research. *Cognitive systems research*, **6**.

Guinan, J.J. Jr. 2010. Cochlear efferent innervation and function. *Current opinion in otolaryngology & head and neck surgery*, **18**(5), 447–453.

Habets, E.A.P. 2006. *Room impulse response generator.* Tech. rept. Technische Universiteit Eindhoven.

Hartmann, M.J., Johnson, N.J., Towal, R.B., & Assad, C. 2003. Mechanical characteristics of rat vibrissae : resonant frequencies and damping in isolated whiskers and in the awake behaving animal. *Journal of neuroscience*, **23**(16), 6510–6519.

Hartmann, W.M. 1997. *Signals, sound and sensation.* American Institute of Physics.

Hartmann, W.M., & Constan, Z.A. 2002. Interaural level differences and the level-meter model. *Journal of the acoustical society of america*, **112**(3), 1037–1045.

Harvey, I., Di Paolo, E., Wood, R., Quinn, M., Tuci, E., & Iridia, E.T. 2005. Evolutionary robotics : A new scientific tool for studying cognition. *Artificial life*, **11**(1-2), 79–98.

Haustein, B.G., & Schrimer, W. 1970. A measuring apparatus for the investigation of the faculty of directional localization. *Hochfrequenztech und elektroakustik*, **79**, 96–101.

Heckmann, M., Rodemann, T., Scholling, B., Joublin, F., & Goerick, C. 2006. Auditory inspired binaural robust sound source localization in echoic and noisy environments. *Pages 368–373 of : Ieee/rsj international conference on intelligent robots and systems*. Citeseer.

Held, R., & Hein, A. 1963. Movement-produced stimulation in the development of visually guided behavior. *Journal of comparative and physiological psychology*, **56**(5), 872–876.

Hofman, P.M. M, Van Riswick, J.G.A., Van Opstal, A.J., Riswick, J G A Van, & Opstal, A J Van. 1998. Relearning sound localization with new ears. *Nature neuroscience*, **1**(5), 417–421.

Horchler, A.D., Reeve, R.E., Webb, B.H., & Quinn, R.D. 2004. Robot phonotaxis in the wild : A biologically inspired approach to outdoor sound localization. *Advanced robotics*, **18**(8), 801 – 816.

Hörnstein, J., Lopes, M., Santos-Victor, J., & Lacerda, F. 2006. Sound localization for humanoid robots-building audio-motor maps based on the HRTF. *Ieee international conference on intelligent robots and systems*, 1170–1176.

Igel, C., & Hüskel, M. 2000. Improving the rprop learning algorithm. *International symposium on neural computation*, 115–121.

Ince, G., Nakadai, K., Rodemann, T., Hasegawa, Y., Tsujino, H., & Imura, J. 2010. A hybrid framework for ego noise cancellation of a robot. *Ieee international conference on robotics and automation*, May, 3623–3628.

Ince, G., Nakamura, K., Asano, F., Nakajima, H., & Nakadai, K. 2011a. Assessment of general applicability of ego noise estimation – applications to automatic speech recognition and sound source localization. *Pages 3517–3522 of : Ieee international conference on robotics and automation*.

Ince, G., Nakadai, K., Rodemann, T., Imura, J., Nakamura, K., & Nakajima, H. 2011b. Assessment of single-channel ego noise estimation methods. *Ieee/rsj international conference on intelligent robots and systems*, 106–111.

Ince, G., Nakadai, K., Rodemann, T., Imura, J., Nakamura, K., & Nakajima, H. 2011c. Incremental learning for ego noise estimation of a robot. *Ieee/rsj international conference on intelligent robots and systems*, 131–136.

Irino, T., & Patterson, R.D. 1997. A time-domain, level-dependent auditory filter : The gammachirp. *Journal of the acoustical society of america*, **101**(1), 412.

Irino, T., & Patterson, R.D. 2001. A compressive gammachirp auditory filter for both physiological and psychophysical data. *Journal of the acoustical society of america*, **109**(5), 2008.

Irino, T., & Patterson, R.D. 2006. A dynamic compressive gammachirp auditory filterbank. *Ieee transactions on audio, speech, and language processing*, **14**(6), 2222–2232.

Ito, A., Kanayama, T., Suzuki, M., & Makino, S. 2005. Internal noise suppression for speech recognition by small robots. *European conference on speech communication and technology*, 2685–2688.

Jeffress, L.A. 1948. A place theory of sound localization. *Journal of comparative and physiological psychology*, **41**(1), 35–39.

Joris, P.X., & Yin, T.C.T. 2007. A matter of time : internal delays in binaural processing. *Trends in neurosciences*, **30**(2), 70–78.

Joris, P.X. X, Smith, P.H. H, & Yin, T.C.T. C. 1998. Coincidence detection in the auditory system : 50 years after Jeffress. *Neuron*, **21**(6), 1235–1238.

Julian, P., Andreou, A.G., Riddle, L., Shamma, S.A., Goldberg, D.H., & Cauwenberghs, G. 2004. A comparative study of sound localization algorithms for energy aware sensor network nodes. *Ieee transactions on circuits and systems*, **51**(4), 640–648.

Kanwal, J.S., & Ehret, G. (eds). 2006. *Behavior and neurodynamics for auditory communication*. Cambridge University Press.

Katsiamis, A.G., Drakakis, E.M., & Lyon, R.F. 2007. Practical gammatone-like filters for auditory processing. *Eurasip journal on audio, speech, and music processing*, **2007**(4), 1–15.

Katsiamis, A.G., Drakakis, E.M., & Lyon, R.F. 2009. A biomimetic, 4.5 uW, 120+ dB, log-domain cochlea channel with AGC. *Ieee journal of solid-state circuits*, **44**(3), 1006–1022.

Kearsley, R.B. 1973. The newborn's response to auditory stimulation : A demonstration of orienting and defensive behavior. *Child development*, **44**(3), 582–590.

Kelly, J.B., & Potash, M. 1986. Directional responses to sounds in young gerbils. *Journal of comparative psychology*, **100**(1), 37–45.

Khalidov, V., Forbes, F., Hansard, M., Arnaud, E., & Horaud, R. 2008. Detection and localization of 3D audio-visual objects using unsupervised clustering. *International conference on multimodal interfaces*, 217–224.

Khalidov, Vasil, Forbes, F., & Horaud, R. 2011. Conjugate mixture models for clustering multimodal data. *Neural computation*, **23**(2), 517–57.

Kidd Jr, G., Arbogast, T.L., Mason, C.R., & Gallun, F.J. 2005. The advantage of knowing where to listen. *Journal of the acoustical society of america*, **118**(6), 3804–3815.

Kim, Y.I., An, S.J., Kil, R.M., & Park, H.M. 2006. Zero-crossing based time-frequency masking for sound segregation. *Neural information processing-letter & review*, **10**(4-6), 125–134.

King, A.J. 2009. Visual influences on auditory spatial learning. *Philosophical transactions of the royal society*, **364**(1), 331–339.

King, AJ, & Middlebrooks, JC. 2011. Cortical representation of auditory space. *Chap. 15, pages 329–341 of :* Winer, J.A., & Schreiner, C.E. (eds), *The auditory cortex.* Springer.

King, A.J., Schnupp, J.W.H., & Doubell, T.P. 2001. The shape of ears to come : dynamic coding of auditory space. *Trends in cognitive sciences*, **5**(6), 261–270.

Knapp, C., & Carter, G. 1976. The generalized correlation method for estimation of time delay. *Ieee transactions on acoustics, speech and signal processing*, **24**, 320–327.

Kneip, L., & Baumann, C. 2008. Binaural model for artificial spatial sound localization based on interaural time delays and movements of the interaural axis. *Journal of the acoustical society of america*, **124**(5), 3108–3019.

Kohlrausch, A., Braasch, J., Kolossa, D., & Blauert, J. 2013. An introduction to binaural processing. *Chap. 1 of :* Blauert, J. (ed), *The technology of binaural listening.* Springer.

Krauzlis, R.J., Basso, M.A., & Wurtz, R.H. 1997. Shared motor error for multiple eye movements. *Science*, **276**(5319), 1693–1695.

Krumbholz, K., Schönwiesner, M., Rübsamen, R., Zilles, K., Fink, G.R., & von Cramon, D.Y. 2005. Hierarchical processing of sound location and motion in the human brainstem and planum temporale. *European journal of neuroscience*, **21**(1), 230–8.

Kumon, M., & Noda, Y. 2011. Active soft pinnae for robots. *International conference on intelligent robots and systems*, 112–117.

Kumon, M., & Uozumi, S. 2011. Binaural localization for a mobile sound source. *Journal of biomechanical science and engineering*, **6**(1), 26–39.

Laflaquière, A. 2013. *Approche sensorimotrice de la perception de l'espace pour la robotique autonome.* Ph.D. thesis, Université Pierre et Marie Curie.

Laflaquière, A., Argentieri, S., Gas, B., & Castillo-Castaneda, E. 2010. Space dimension perception from the multimodal sensorimotor flow of a naive robotic agent. *Ieee/rsj international conference on intelligent robots and systems*, 1520–1525.

Lang, P.J., Bradley, M.M., & Cuthbert, B.N. 1990. Emotion, attention, and the startle reflex. *Psychological review*, **97**(3), 377–395.

Larcher, V., & Jot, J.M. 1997. Techniques d'interpolation de filtres audio-numeriques. Application a la reproduction spatiale des sons sur écouteurs. *Congrès de la société française d'acoustique.*

Lazzaro, J., & Mead, C.A. 1989. A silicon model of auditory localization. *Neural computation*, **1**, 47–57.

Lee, J.A., & Verleysen, M. 2010. *Nonlinear dimensionality reduction.* Springer.

Lee, S., Hwang, S., & Park, Y. 2008. Sound source localization in median plane using artificial ear. *International conference on control, automation and systems*, Oct., 246–250.

Leong, M.P., Jin, C.T., & Leong, P.H.W. 2003. An FPGA-based electronic cochlea. *Eurasip journal on advances in signal processing*, **7**, 629–638.

Lewicki, M.S. 2002. Efficient coding of natural sounds. *Nature neuroscience*, **5**(4), 356–63.

Licklider, J.C.R. C. R., & Pollack, I. 1948. Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. *Journal of the acoustical society of america*, **20**(1), 42–51.

Litovsky, R.Y., Colburn, H.S., Yost, W.A., & Guzman, S.J. 1999. The precedence effect. *Journal of the acoustical society of america*, **106**(4), 1633–1654.

Liu, J., Erwin, H., Wermter, S., & Elsaid, M. 2008. A biologically inspired spiking neural network for sound localisation by the inferior colliculus. *Artificial neural networks - lecture notes in computer science*, **5164**, 396–405.

Liu, J., Perez-Gonzalez, D., Rees, A., Erwin, H., & Wermter, S. 2009. Multiple sound source localisation in reverberant environments inspired by the auditory midbrain. *International confenrence on artificial neural networks*, 208–217.

Liu, S.C., van Schaik, A., Minch, B.A., & Delbruck, T. 2010. Event-based 64-channel binaural silicon cochlea with Q enhancement mechanisms. *Ieee international symposium on circuits and systems*, 2027–2030.

Loomis, J.M., Hebert, C, & Cicinelli, J.G. 1990. Active localization of virtual sounds. *Journal of the acoustical society of america*, **88**(4), 1757–64.

Lopez-Poveda, E.A., & Meddis, R. 1996. A physical model of sound diffraction and reflections in the human concha. *Journal of the acoustical society of america*, **100**(5), 3248–59.

Lu, Y.C., Cooke, M., & Christensen, H. 2007. Active binaural distance estimation for dynamic sources. *Conference of the international speech communication association*, 1–4.

Lungarella, M., Metta, G., Pfeifer, R., & Sandini, G. 2003. Developmental robotics : a survey. *Connection science*, **15**(4), 151–190.

Lyon, R.F. 1982. A computational model of filtering, detection, and compression in the cochlea. *Ieee international conference on acoustics, speech, and signal processing*, **7**, 1282–1285.

Lyon, R.F. 1983. A computational model of binaural localization and separation. *Ieee international conference on on acoustics, speech, and signal processing*, **8**, 1148–1151.

Lyon, R.F. 1990. Automatic gain control in cochlear mechanics. *The mechanics and biophysics of hearing - lecture notes in biomathematics*, 395–402.

Lyon, R.F. 1996. *The all-pole gammatone filter and auditory models.* Tech. rept. Apple Computer.

Lyon, R.F. 2010. Machine hearing : an emerging field. *Ieee signal processing magazine*, 131–136.

Lyon, R.F., Katsiamis, A.G., & Drakakis, E.M. 2010. History and future of auditory filter models. *Ieee international symposium on circuits and systems*, 3809–3812.

Ma, N. 2006. *An efficient implementation of gammatone filters.*

Ma, N., Green, P., Barker, J., & Coy, A. 2007. Exploiting correlogram structure for robust speech recognition with multiple speech sources. *Speech communication*, **49**(12), 874–891.

Macpherson, E.A., & Middlebrooks, J.C. 2002. Listener weighting of cues for lateral angle : The duplex theory of sound localization revisited. *Journal of the acoustical society of america*, **111**(5), 2219–2236.

Manoonpong, P., Pasemann, F., Fischer, J., & Roth, H. 2005. Neural processing of auditory signals and modular neural control for sound tropism of walking machines. *International journal of advanced robotic systems*, **2**(3), 223–234.

May, T., van de Par, S., & Kohlrausch, A. 2011. A probabilistic model for robust localization based on a binaural auditory front-end. *Ieee transactions on audio, speech, and language processing*, **19**(1), 1–13.

McAlpine, D. 2003. Sound localization and delay lines – do mammals fit the model ? *Trends in neurosciences*, **26**(7), 347–350.

McAlpine, David. 2005. Creating a sense of auditory space. *Journal of physiology*, **566**(1), 21–8.

McDermott, J.H. 2009. The cocktail party problem. *Current biology*, **19**(22), R1024–7.

McDermott, J.H., Wrobleski, D., & Oxenham, A.J. 2011. Recovering sound sources from embedded repetition. *Proceedings of the national academy of sciences of the united states of america*, **108**(3).

Meddis, R. 1986. Simulation of mechanical to neural transduction in the auditory receptor. *Journal of the acoustical society of america*, **79**(3), 702–711.

Meddis, R., Hewitt, M.J., & Shackleton, T.M. 1990. Implementation details of a computation model of the inner hair-cell auditory-nerve synapse. *Journal of the acoustical society of america*, **87**(4), 1813–1816.

Mershon, D.H., & Bowers, J.N. 1979. Absolute and relative cues for the auditory perception of egocentric distance. *Perception*, **8**(3), 311–22.

Metta, G. 2000. *Babyrobot, a study on sensori-motor development.* Ph.D. thesis.

Meyer, J.A., Guillot, A, Girard, B, Khamassi, M, & P. 2005. The Psikharpax project : Towards building an artificial rat. *Robotics and automation*, **50**, 211–223.

Moddemeijer, R. 1988. An information theoretical delay estimator. *Ninth symosium on information theory in the benelux*, 121–128.

Moore, B.C.J. 1997. *An introduction to the psychology of hearing.* London Academic Press.

Moore, B.C.J. 2003. Coding of sounds in the auditory system and its relevance to signal processing and coding in cochlear implants. *Otology and neurotology*, **24**(2), 243–54.

Moravec, HP. 1988. *Mind Children.* Cambridge University Press.

Muir, D.W., Clifton, R.K., & Clarkson, M.G. 1989. The development of a human auditory localization response : A U-shaped function. *Canadian journal of psychology*, **43**(2), 199–216.

Murray, John C, Erwin, Harry, & Wermter, Stefan. 2004. Robotic sound-source localization and tracking using interaural time difference and cross-correlation. *Proceedings of neurobotics workshop.*

Musicant, A.D., & Butler, R.A. 1984. The influence of pinnae-based spectral cues on sound localization. *Journal of the acoustical society of america*, **75**(4), 1195–200.

Naguib, M. 2001. Estimating the distance to a source of sound : mechanisms and adaptations for long-range communication. *Animal behaviour*, **62**(5), 825–837.

Nakadai, K., Lourens, T., Okuno, H.G., & Kitano, H. 2000a. Active audition for humanoid. *National conference on artificial intelligence.*

Nakadai, K., Matsui, T., Okuno, H.G., & Kitano, H. 2000b. Active audition system and humanoid exterior design. *Ieee international conference on intelligent robots and systems*, 1453–1461.

Nakadai, K., Okuno, H.G., & Kitano, H. 2000c. Humanoid active audition system improved by the cover acoustics. *Pricai 2000 topics in artificial intelligence - lecture notes in computer science*, **1886**, 544–554.

Nakajima, H., Kikuchi, K., Daigo, T., Kaneda, Y., Nakadai, K., & Hasegawa, Y. 2009. Real-time sound source orientation estimation using a 96 channel microphone array. *Ieee international conference on intelligent robots and systems*, 676–683.

Neimark, M.A., Andermann, M.L., Hopfield, J.J., & Moore, C.I. 2003. Vibrissa resonance as a transduction mechanism for tactile encoding. *Journal of neuroscience*, **23**(16), 6499–6509.

Nguyen, M.D., Inaba, A., Suzuki, A., & Takahashi, H. 2010. Sound direction sensor with an acoustic channel. *Ieee international conference on micro electro mechanical systems*, Jan., 655–658.

N'Guyen, S. 2010. *Mise au point du système vibrissal du robot-rat Psikharpax et contribution à la fusion de ses capacités visuelle, auditive et tactile.* Ph.D. thesis, Université Pierre et Marie Curie.

N'Guyen, S., Pirim, Patrick, & Meyer, J.A. 2009. Elastomer-based tactile sensor array for the artificial rat Psikharpax. *In : Isef 2009 - xiv international symposium on fields in mechatronics, electrical and electronic engineering.*

N'Guyen, S., Pirim, P., & Meyer, J.A. 2011a. Tactile texture discrimination in the robotrat Psikharpax. *Biomedical engineering systems and technologies - communications in computer and information science*, **127**(2003), 252–265.

N'Guyen, S., Pirim, P., Meyer, J.A., & N'Guyen, Steve. 2011b. Texture discrimination with artificial whiskers in the robot-rat Psikharpax. *Biomedical engineering systems and technologies - communications in computer and information science*, **127**, 252–266.

Nishimura, M., Nakano, M., Nakadai, K., Tsujino, H., & Ishizuka, M. 2006. Speech recognition for a robot under its motor noises by selective application of missing feature theory and MLLR. *Tutorial and research workshop on statistical and perceptual audition.*

Nissen, S. 2003. *Implementation of a fast artificial neural network library.* Tech. rept. Department of Computer Science, University of Copenhagen.

Noë, A., & O'Regan, J.K. 2000. Perception, attention, and the grand illusion. *Psyche*, **6**(15).

Oertel, D., Fay, R.R., & Popper, A.N. (eds). 2001. *Integrative functions in the mammalian auditory pathway.* Springer Handbook of Auditory Research.

Oertel, Donata, & Young, E.D. 2004. What's a cerebellar circuit doing in the auditory system ? *Trends in neurosciences*, **27**(2), 104–10.

O'Regan, J.K. 2011. *Why red doesn't sound like a bell : Understanding the feel of consciousness.* Oxford University Press.

O'Regan, J.K., & Noë, A. 2001. A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, **24**(5), 939–1031.

Oxenham, A.J. 2001. Forward masking : Adaptation or integration ? *Journal of the acoustical society of america*, **109**(2), 732–741.

Palmer, A.R., Hall, D.A., Sumner, C., Barrett, D.J.K., Jones, S., Nakamoto, K., & Moore, D.R. 2007. Some investigations into non-passive listening. *Hearing research*, **229**(1-2), 148–57.

Park, H.M. 2006. Spatial separation of speech signals using continuously-variable masks estimated from comparisons of zero crossings. *Ieee international conference on acoustics, speech and signal processing*, **4**, 1165–1168.

Park, Y., & Hwang, S. 2007. Artificial Robot Ear Design for Sound Direction Estimation. *Ieee international symposium on robot and human interactive communication*, **59**(3-4), 405–409.

Patterson, R.D., Nimmo-Smith, I., Holdsworth, J., & Rice, P. 1987. *An efficient auditory filterbank based on the gammatone function.* Tech. rept. December. Cambridge Electronic Design.

Patterson, R.D., Robinson, K., Holdsworth, J., DMcKeown, D., Zhang, C., & Allerhand, M. 1992. Complex sounds and auditory images. *International symposium on hearing*, 429–446.

Patterson, R.D., Allerhand, M.H., & Giguère, C. 1995. Time-domain modeling of peripheral auditory processing : A modular architecture and a software platform. *Journal of acoustical society of america*, **98**(4), 1890–1894.

Patterson, R.D., Unoki, M., & Irino, T. 2003. Extending the domain of center frequencies for the compressive gammachirp auditory filter. *Journal of the acoustical society of america*, **114**(3), 1529.

Pearson, M.J. J., Pipe, A.G. G., Melhuish, C., Mitchison, B., Prescott, T.J. J., & Mitchinson, B. 2007. Whiskerbot : A robotic active touch system modeled on the rat whisker sensory system. *Adaptive behavior*, **15**(3), 223–240.

Peremans, H., & Muller, R. 2000. A comprehensive robotic model for neural & acoustic signal processing in bats. *Ieee/embs international conference on neural engineering*, 458–461.

Peremans, H., & Reijniers, Jonas. 2005. The CIRCE head : a biomimetic sonar system. *International conference on artificial neural networks*, 283–288.

Perrett, S, & Noble, W. 1997. The effect of head rotations on vertical plane sound localization. *Journal of the acoustical society of america*, **102**(4), 2325–32.

Petersen, C.H. 2007. The functional organization of the barrel cortex. *Neuron*, **56**(2), 339–355.

Peterson, L.E. 2009. K-nearest neighbor. *Scholarpedia*, **4**(2), 1883.

Philipona, D, O'Regan, J.K., & Nadal, J-P J.P. J.-P. 2003. Is there something out there ? Inferring space from sensorimotor dependencies. *Neural computation*, **15**(9), 2029–49.

Philipona, D, O'Regan, JK, Nadal, J.P., & Coenen, O. 2004. Perception of the structure of the physical world using unknown multimodal sensors and effectors. *Advances in neural information processing systems*, **16**, 945–952.

Piéron, H. 1922. L'orientation auditive latérale. *L'année psychologique*, **23**(1), 186–213.

Pinho, C., Ferreira, J.F., Bessière, P., & Dias, J. 2008. A Bayesian binaural system for 3D sound-source localisation. *International conference on cognitive systems*, 109–114.

Pitti, A., Blanchard, A., Cardinaux, M., & Gaussier, P. 2012a. Distinct mechanisms for multimodal integration and unimodal representation in spatial development. *Ieee international conference on development and learning*, 1–6.

Pitti, A., Blanchard, A., Cardinaux, M., & Gaussier, P. 2012b. Gain-field modulation mechanism in multimodal networks for spatial perception. *Ieee international conference on humanoid robots*, 297–302.

Plack, C.J., & Oxenham, A.J. 1998. Basilar-membrane nonlinearity and the growth of forward masking. *Journal of the acoustical society of america*, **103**(3), 1598–608.

Plack, C.J., Oxenham, A.J., & Drga, V. 2002. Linear and nonlinear processes in temporal masking. *Acta acustica united with acustica*, **88**, 348–358.

Poincaré, H. 1895. L'espace et la géométrie. *Revue de métaphysique et de morale*, **3ème année**(6), 631–646.

Populin, L.C., & Yin, T.C.T. 1998. Pinna movements of the cat during sound localization. *Journal of neuroscience*, **18**(11), 4233–43.

Portello, A., Danes, P., & Argentieri, S. 2011. Acoustic models and Kalman filtering strategies for active binaural sound localization. *Pages 137–142 of : Ieee/rsj international conference on intelligent robots and systems*. IEEE.

Preibisch-Effenberger, R. 1966. *The human faculty of sound localization and its audiometric application to clinicla diagnostics*. Ph.D. thesis.

Prescott, T.J., Pearson, M.J., Mitchinson, B., Sullivan, J.C.W., & Pipe, A.G. 2009. Whisking with robots. *Ieee robotics and automation magazine*, **16**(September), 42–50.

Pressnitzer, D., & Gnansia, D. 2005. Real-time auditory models. *International computer music conference*, 295–298.

Rañó, I. 2007. On taxis for control and its qualitative solution on mobile robots. *Proceedings of the intelligent autonomous vehicles*.

Rañó, I. 2012. A systematic analysis of the Braitenberg vehicle 2b for point-like stimulus sources. *Bioinspiration & biomimetics*, **7**(3).

Raspaud, M., Viste, H., & Evangelista, G. 2010. Binaural source localization by joint estimation of ILD and ITD. *Ieee transactions on audio, speech, and language processing*, **18**(1), 68–77.

Rayleigh, Lord. 1907. On our perception of sound direction. *Philosophical magazine*, **13**(74), 214–232.

Reeve, R.E., & Webb, B.H. 2003. New neural circuits for robot phonotaxis. *Philosophical transactions of the royal society of london*, **361**, 2245–2266.

Reeve, R.E., Webb, B.H., Horchler, A.D., Indiveri, G., & Quinn, R. 2005. New technologies for testing a model of cricket phonotaxis on an outdoor robot. *Robotics and autonomous*, **51**, 41–54.

Reid, G.L., & Milios, E. 2003 (Jan.). *Active stereo sound localization*. Tech. rept. 1. York University.

Rice, J.J., May, B.J., Spirou, G.A., & Young, E.D. 1992. Pinna-based spectral cues for sound localization in cat. *Hearing research*, **58**(2), 132–152.

Robles, L., & Delano, P.H. 2008. Efferent system. *In :* Dallos, P., & Oertel, D. (eds), *The senses : a comprehensive reference*. Elsevier.

Robles, L., & Ruggero, M.A. 2001. Mechanics of the mammalian cochlea. *Physiological reviews*, **81**(3), 1305–1352.

Rodemann, T. 2010. A study on distance estimation in binaural sound localization. *Ieee/rsj international conference on intelligent robots and systems*, 425–430.

Rodemann, T. 2011. Spectral cues to source position in robots with arbitrary ear shapes. *International conference on advanced robotics*, June, 453–458.

Rodemann, T., Heckmann, M., Joublin, F., Goerick, C., & Scholling, B. 2006. Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping. *Ieee/rsj international conference on intelligent robots and systems*, Oct., 860–865.

Rodemann, T., Karova, K., Joublin, F., & Goerick, C. 2007. Purely auditory online-adaptation of auditory-motor maps. *Ieee/rsj international conference on intelligent robots and systems*, Oct., 2015–2020.

Rodemann, T., Ince, G., Joublin, F., & Goerick, C. 2008. Using binaural and spectral cues for azimuth and elevation localization. *Ieee/rsj international conference on intelligent robots and systems*, 2185 – 2190.

Roffler, S.K., & Butler, R.A. 1968. Factors that influence the localization of sound in the vertical plane. *Journal of the acoustical society of america*, **43**, 1259–1288.

Roman, N., Wang, D.L., & Brown, G.J. 2003. Speech segregation based on sound localization. *Journal of the acoustical society*, **114**(4), 2236–2252.

Roschin, V.Y., Frolov, A.A., Burnod, Y., & Maier, M.A. 2011. A neural network model for the acquisition of a spatial body scheme through sensorimotor interaction. *Neural computation*, **23**(7), 1821–34.

Rucci, M., & Wray, J. 1999. Binaural cross-correlation and auditory localization in the barn owl : a theoretical study. *Neural networks*, **12**(1), 31–42.

Rucci, M., Edelman, G.M., & Wray, J. 1999. Adaptation of orienting behavior : from the barn owl to a robotic system. *Ieee transactions on robotics and automation*, **15**(1), 96–110.

Rucci, M., Wray, J., & Edelman, G.M. 2000. Robust localization of auditory and visual targets in a robotic barn owl. *Robotics and autonomous systems*, **30**(1-2), 181–193.

Satarzadeh, P., Algazi, V.R., & Duda, R.O. 2007. Physical and filter pinna models based on anthropometry. *Pages 5–8 of : Convention of the audio engineering society*. Citeseer.

Schauer, C., & Gross, H.M. 2003. A computational model of early auditory-visual integration. *Pattern recognition - lecture notes in computer science*, **2781**, 362–369.

Schwartz, O., & Simoncelli, E.P. 2001. Natural signal statistics and sensory gain control. *Nature neuroscience*, **4**(8), 819–25.

Searle, C.L., Braida, L.D., Cuddy, D.R., & Davis, M.F. 1975. Binaural pinna disparity : another auditory localization cue. *Journal of the acoustical society of america*, **57**(2), 448–55.

Shamma, S.A. 1989. Stereausis : Binaural processing without neural delays. *Journal of the acoustical society of america*, **86**(3), 989.

Shamma, S.A., & Micheyl, C. 2010. Behind the scenes of auditory perception. *Current opinion in neurobiology*, **20**(Apr.), 1–6.

Shamma, S.A., Elhilali, M., & Micheyl, C. 2011. Temporal coherence and attention in auditory scene analysis. *Trends in neurosciences*, **34**(3), 114–123.

Shaw, E.A.G., & Teranishi, R. 1968. Sound pressure generated in an external-ear replica and real human ears by a nearby point source. *Journal of the acoustical society of america*, **44**(1), 240–9.

Shaw, WT. 2006. Sampling Student's T distribution-use of the inverse cumulative distribution function. *Journal of computational finance*, **9**(4), 37–73.

Shepard, D. 1968. A two-dimensional interpolation function for irregularly-spaced data. *Acm national conference*, 517–524.

Shimoda, T., Nakashima, T., Kumon, M., Kohzawa, R., Mizumoto, I., & Iwai, Z. 2007. Sound Localization of Elevation using Pinnae for Auditory Robots. *Chap. 24, pages 421–438 of :* Grimm, M., & Kroschel, K. (eds), *Robust speech recognition and understanding.* I-Tech.

Shinn-Cunningham, B.G., Santarelli, S., & Kopco, N. 2000. Tori of confusion : binaural localization cues for sources within reach of a listener. *Journal of the acoustical society of america*, **107**(3), 1627–36.

Sinyor, A., & Laszlo, C.A. 1973. Acoustic behavior of the outer ear of the guinea pig and the influence. *Journal of the acoustical society of america*, **54**(4), 916–921.

Slaney, M. 1988. *Lyon's cochlear model.* Tech. rept. Apple Computer.

Slaney, M. 1993. *An efficient implementation of the Patterson-Holdsworth auditory filter bank.* Tech. rept. Apple Computer.

Slaney, M. 1998. *Auditory Toolbox.* Tech. rept. Interval Research Corporation.

Slaney, Malcolm. 1997. A critique of pure audition. *Chap. 3 of :* Rosenthal, D., & Okuno, H.G. (eds), *Computaional auditory scene analysis.* Lawrence Erlbaum Associates.

Smith, Evan C E.C., & Lewicki, Michael S M.S. 2006. Efficient auditory coding. *Nature*, **439**(7079), 978–982.

Smith, L.S. 1992. Using IIDs to estimate sound source direction. *International conference on simulaiton of adaptive behavior : from animals to animats*, 60–61.

Smith, L.S., & Collins, S. 2007. Determining ITDs using two microphones on a flat panel during onset intervals with a biologically inspired spike-based technique. *Ieee transactions on audio, speech, and language processing*, **15**(8), 2278–2286.

Speigle, J.M., & Loomis, J.M. 2002. Auditory distance perception by translating observers. *Ieee symposium on research frontiers in virtual reality*, 92–99.

Stangor, C. 2010. *Introduction to psychology.* Flatworld Knowledge.

Stein, B.E., & Meredith, M.A. 1993. *The merging of the senses.* MIT Press.

Stern, R.M., Brown, G.J., & DeLiang, W. 2006. Binaural Sound Localization. *Chap. 5 of :* DeLiang, W., & Brown, G.J. (eds), *Computaional auditory scene analysis : Principles, algorithms and applications.* John Wiley & Sons Inc.

Stone, C.J. 1977. Consistent nonparametric regression. *The annuals of statistics*, **5**(4), 595–620.

Sumner, Christian J. C.J., Lopez-Poveda, Enrique a. E.A., O'Mard, L.P. Lowel P., & Meddis, Ray. 2002. A revised model of the inner-hair cell and auditory-nerve complex. *Journal of the acoustical society of america*, **111**(5), 2178.

Sumner, C.J., Lopez-Poveda, E.A., O'Mard, L.P., & Meddis, R. 2003. Adaptation in a revised inner-hair cell model. *Journal of the acoustical society of america*, **113**(2), 893–901.

Trifa, V.M. Vlad M., Koene, Ansgar, Moren, Jan, & Cheng, Gordon. 2007. Real-time acoustic source localization in noisy environments for human-robot multimodal interaction. *Ieee international symposium on robot and human interactive communication*, 393–398.

Van der Maaten, L.J., Postma, E.O., & Van den Herik, H.J. 2009. Dimensionality reduction : a comparative review. *Tilburg university technical report.*

van Schaik, A., & Shamma, S.A. 2004. A neuromorphic sound localizer for a smart MEMS system. *Analog integrated circuits and signal processing*, **39**(3), 267–273.

Van Trees, H.L. 2002. *Optimum array processing.* Wiley-Interscience.

Vincent, S.B. 1912. The function of the vibrissae in the behavior of the white rat. *Behavior monographs*, **1**(5), 81.

von Békésy, G. 1949. The moon illusion and similar auditory phenomena. *American journal of psychology*, **62**(4), 540–552.

Walker, V.A., Peremans, H., & Hallam, J.C.T. 1998. One tone, two ears, three dimensions : A robotic investigation of pinnae movements used by rhinolophid and hipposiderid bats. *Journal of the acoustical society of america*, **104**, 569–579.

Wallach, H. 1940. The role of head movements and vestibular and visual cues in sound localization. *Journal of experimental psychology*, **27**(4), 339–368.

Wallach, H., Newman, E.B., & Rosenzweig, M.R. 1949. The precedence effect in sound localization. *American journal of psychology*, **62**(3), 315–336.

Wang, J., Neskovic, P., & Cooper, L.N. 2007. Improving nearest neighbor rule with a simple adaptive distance measure. *Pattern recognition letters*, **28**(2), 207–213.

Warren, Richard M. 2008. *Auditory perception : an analysis and synthesis.* 3rd edn. Cambridge University Press.

Webb, B.H. 2001. Can robots make good models of biological behaviour ? *Behavioral and brain sciences*, **24**, 1033–1094.

Webb, B.H., & Scutt, T. 2000. A simple latency-dependent spiking-neuron model of cricket phonotaxis. *Biological cybernetics*, **82**, 247–269.

Weiss, R.J., Mandel, M.I., & Ellis, D.P.W. 2011. Combining localization cues and source model constraints for binaural source separation. *Speech communication*, **53**(5), 606–621.

Werner, L.A. 2008. Human auditory development. *In :* Dallos, P., & Oertel, D. (eds), *The senses : a comprehensive reference.* Elsevier.

Wiener, F. 1947. On the diffraction of a progressive wave by the human head. *Journal of the acoustical society of america*, **19**, 143–146.

Wierstorf, H., Geier, M., Raake, A., & Spors, S. 2011. A free database of head-related impulse response measurements in the horizontal plane with multiple distances. *In : 130th convention of the audio engineering society.*

Wilson, B.S., & Dorman, M.F. 2008. Cochlear implants : a remarkable past and a brilliant future. *Hearing research*, **242**(1-2), 3–21.

Wilson, B.S., Lawson, D.T., & Zerbi, M. 1994 (Nov.). *Speech processors for auditory prostheses.* Tech. rept. 9. Center for Auditory Prosthesis Research, Research Triangle Institute.

Winer, J.A., & Schreiner, C.E. 2011. *The auditory cortex.* Springer.

Wrigley, S.N., & Brown, G.J. 2007. Binaural speech separation using recurrent timing neural networks for joint F0-localisation estimation. *international conference on machine learning for multimodal interaction*, 271–282.

Yamamoto, S., Nakadai, K., Nakano, M., Tsujino, H., Valin, J.M., Komatani, K., Ogata, T., & Okuno, H.G. 2006. Real-time robot audition system that recognizes simultaneous speech in the real world. *Ieee/rsj international conference on intelligent robots and systems*, Oct., 5333–5338.

Yan, R., Rodemann, T., & Wrede, B. 2011. Learning of audiovisual integration. *Ieee international conference on development and learning*, **2**, 1–7.

Yin, T.C.T. 2002. Neural mechanisms of encoding binaural localization cues in the auditory brainstem. *Chap. 4, pages 99–159 of :* Fay, R.R., & Popper, A.N. (eds), *Integrative functions in the mammalian auditory pathway.* Springer Handbook of Auditory Research.

Yost, W.A., Popper, A.N., & Fay, R.R. (eds). 2008. *Auditory perception of sound sources.* Springer Handbook of Auditory Research, vol. 16, no. 5. Springer Handbook of Auditory Research.

Young, P.T. 1931. The role of head movements in auditory localization. *Journal of experimental psychology*, **14**(2), 95–124.

Youssef, K., Argentieri, S., & Zarader, J.L. 2010. From monaural to binaural speaker recognition for humanoid robots. *Ieee-ras international conference on humanoid robots*, Dec., 580–586.

Youssef, K., Argentieri, S., & Zarader, J.L. 2011. Multimodal sound localization for humanoid robots based on visio-auditive learning. *International conference on robotics and biomimetics*, Dec., 2517–2522.

Youssef, K., Argentieri, S., & Zarader, J.L. 2012a. A binaural sound source localization method using auditive cues and vision. *Ieee international conference on acoustics, speech and signal processing.*

Youssef, K., Argentieri, S., & Zarader, J.-L. 2012b. Towards a systematic study of binaural cues. *Pages 1004–1009 of : 2012 ieee/rsj international conference on intelligent robots and systems.* Ieee.

Youssef, K., Argentieri, S., & Zarader, J.L. 2013. A learning-based approach to robust binaural sound localization. *International conference on intelligent robots and systems.*

Zahorik, P. 1996. *Auditory distance perception : A literature review.* Tech. rept. University of Wisconsin.

Zahorik, P. 2002. Auditory display of sound source distance. *Pages 326–332 of : International conference on auditory display.* Citeseer.

Zakarauskas, P., & Cynader, M.S. 1993. A computational theory of spectral cue localization. *Journal of the acoustical society of america*, **94**(3), 1323.

Zatorre, R.J., Bouffard, M., Ahad, P., & Belin, P. 2002. Where is 'where' in the human auditory cortex ? *Nature neuroscience*, **5**(9), 905–9.

Zhang, Z., & Zha, H. 2002. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *Arxiv preprint cs/0212008.*