Modulation de Mouvements de Tête pour l'Analyse Multimodale d'un Environnement Inconnu

Benjamin Cohen-Lhyver

Université Pierre et Marie Curie

Ecole Doctorale SMAER
Sciences Mécaniques, Acoustique, Electronique et Robotique de Paris

Institut des Systèmes Intelligents et de Robotique

Modulation de Mouvements de Tête pour l'Analyse Multimodale d'un Environnement Inconnu

Benjamin Cohen-Lhyver

Sous la supervision du Pr. Bruno Gas

et encadré par Sylvain Argentieri

Thèse soutenue publiquement le mardi 19 septembre 2017,
devant le jury composé de :

QUOY	Mathias	Professeur des Universités	Rapporteur
LOPES	Manuel	Chargé de Recherche, HDR	Rapporteur
BLAUERT	Jens	Professeur émérite	Examinateur
ROEBEL	Axel	Directeur de Recherche	Examinateur
DELEFORGE	Antoine	Chargé de Recherche	Examinateur
GAS	Bruno	Professeur des Universités	Directeur de thèse
ARGENTIERI	Sylvain	Maître de Conférences	Encadrant de thèse

Pour obtenir le grade de Docteur
Spécialité Sciences de l'Ingénieur
Université Pierre et Marie Curie, 1er septembre 2017

Résumé

L'exploration d'un environnement inconnu par un robot mobile est un vaste domaine de recherche visant à comprendre et implémenter des modèles d'exploration efficaces, rapides et pertinents. Cependant, depuis les années 80, l'exploration ne s'est plus contentée de la seule détermination de la topographie d'un espace : à la composante spatiale a été couplée une composante sémantique du monde exploré. En effet, en addition aux caractéristiques physiques de l'environnement — murs, obstacles, chemins empruntables ou non, entrées et sorties — permettant au robot de se créer une représentation interne du monde grâce à laquelle il peut s'y déplacer, existent des caractéristiques dynamiques telles que l'apparition d'événements audiovisuels. Ces événements sont d'une grande importance en cela qu'ils peuvent moduler le comportement du robot en fonction de leur localisation dans l'espace — aspect topographique — et de l'information qu'ils portent — aspect sémantique. Bien qu'imprédictibles par nature (puisque l'environnement est inconnu) tous ces événements ne sont pas d'égale importance : certains peuvent porter une information utile au robot et à sa tâche d'exploration, d'autres non.

Suivant les travaux sur les motivations intrinsèques à explorer un environnement inconnu et puisant son inspiration de phénomènes neurologiques, ce travail de thèse a consisté en l'élaboration du modèle Head Turning Modulation (HTM) visant à donner à un robot doté de mouvements de tête la capacité de déterminer l'importance relative de l'apparition d'un événement audiovisuel dans un environnement inconnu en cours d'exploration. Cette « importance » a été formalisée sous la forme de la notion de *Congruence* s'inspirant principalement (i) de l'entropie de Shannon, (ii) du phénomène de *Mismatch Negativity* et (iii) de la *Reverse Hierarchy Theory*. Le modèle HTM, créé dans le cadre du projet européen Two!Ears, est un paradigme d'apprentissage basé sur (i) une *auto-supervision* (le robot décide lorsqu'il est nécessaire d'apprendre ou non), (ii) une contrainte de *temps réel* (le robot apprend et réagit aussitôt que des données sont perçues), et (iii) une absence de données *a priori* sur l'environnement (il n'existe pas de « vérité » à apprendre, seulement la réalité perçue de l'environnement à explorer). Ce modèle, intégré à l'ensemble du *framework* Two!Ears, a été entièrement porté sur un robot mobile pourvu d'une vision binoculaire et d'une audition binaurale. Le modèle HTM couple ainsi une approche montante traditionnelle d'analyse des signaux perceptifs (extractions de caractéristiques, reconnaissance visuelle ou auditive, etc.) à une approche descendante permettant, via la génération d'une action motrice, de comprendre et interpréter l'environnement audiovisuel du robot. Cette approche *bottom-up/top-down* active est ainsi exploitée pour moduler les mouvements de tête d'un robot humanoïde et étudier l'impact de la Congruence sur ces mouvements. Le système a été évalué *via* des simulations réalistes, ainsi que dans des conditions réelles, sur les deux plateformes robotiques du projet Two!Ears.

Abstract

The exploration of an unknown environement by a mobile robot is a vast research domain aiming at understanding and implementing efficient, fast and relevant exploration models. However, since the 80s, exploration is no longer restricted to the sole determination of topography a space : to the spatial component has been coupled a semantic one of the explored world. Indeed, in addition to the physical characteristics of the environment — walls, obstacles, usable paths or not, entrances and exits — allowing the robot to create its own internal representation of the world through which it can move in it, exist dynamic components such as the apparition of audiovisual events. These events are of high importance for they can modulate the robot's behavior through their location in space — topographic aspect — and the information they carry — semantic aspect. Although impredictible by nature (since the environment is unknown) all these events are not of equal importance : some carry valuable information for the robot's exploration task, some don't.

Following the work on intrinsic motivations to explore an unknown environment, and being rooted in neurological phenomenons, this thesis work consisted in the elaboration of the Head Turning Modulation (HTM) model aiming at giving to a robot capable of head movements, the ability to determine the relative importance of the apparition of an audioivsual event. This "importance" has been formalized through the notion of *Congruence* which is mainly inspired from (i) Shannon's entropy, (ii) the Mismatch Negativity phenomenon, and (iii) the Reverse Hierarchy Theory. The HTM model, created within the Two!Ears european project, is a learning paradigm based on (i) an *auto-supervision* (the robot decides when it is necessary or not to learn), (ii) a *real-time* constraint (the robot learns and reacts as soon as data is perceived), and (iii) an absence of prior knowledge about the environment (there is no "truth" to learn, only the reality of the environment to explore). This model, integrated in the overal Two!Ears framework, has been entirely implemented in a mobile robot with binocular vision and binaural audition. The HTM model thus gather the traditional approach of ascending analysis of perceived signals (extraction of caracteristics, visual or audio recognition etc.) to a descending approach that enables, via motor actions generation in order to deal with perception deficiency (such as visual occlusion), to understand and interprete the audiovisual environment of the robot. This bottom-up/top-down active approach is then exploited to modulate the head movements of a humanoid robot and to study the impact of the Congruence on these movements. The system has been evaluated via realistic simulations, and in real conditions, on the two robotic platforms of the Two!Ears project.

Remerciements

Cette thèse est l'aboutissement d'un long parcours durant lequel de nombreuses émotions et sentiments se sont succédés, parfois chevauchés. Même s'il s'agit d'un travail éminemment solitaire, il n'aurait pas été possible, ou alors avec une finalité différente, sans toutes les personnes qui m'ont entouré, supporté, encouragé et aidé. Que ce soit par votre soutien direct et chaleureux, par vos preuves d'amitié et d'affection, mais aussi par votre capacité à accepter que je place ce travail avant le reste, je vous remercie du fond du cœur. Je vous remercie d'avoir enduré la litanie que j'ai répétée durant ces presque quatre ans : « j'peux pas, j'ai du taf ». Je vous remercie d'avoir supporté mes absences répétées à nos soirées, nos rendez-vous, nos sorties ou nos projets de vacances. Je vous remercie d'avoir compris mes sautes d'humeurs, mes moments les plus durs et mes baisses de régime. Votre compréhension, et le fait que vous ayez accepté que je mette au premier plan de ma vie ce travail de thèse, m'a également permis d'avancer correctement et d'avoir le temps et les ressources dont j'avais besoin pour l'accomplir. J'ai essayé de faire le maximum pour qu'il ne phagocyte pas trop les liens qui tissent nos différentes relations mais, malgré tous mes efforts, je sais que vous avez dû composer pendant toutes ces années avec une entité hybride aliénante : « moi + ma thèse ».

Mais cette thèse est sans aucun doute une des choses dont je suis le plus fier, accomplissement d'une promesse que je m'étais faite lorsque j'avais dix-sept ans, alors que je fréquentais (déjà) de nombreux cafés parisiens au lieu d'aller au lycée, à une époque où l'on pouvait encore fumer en prenant un café au comptoir — imaginez ma délectation. Ceux qui me connaissent et qui m'ont entouré durant ces quatre ans savent la difficulté avec laquelle j'ai vécu ce doctorat mais aussi toutes les réussites que j'ai connues. C'est ainsi que vous m'avez entendu vouloir tout arrêter environ une fois par semaine, pester contre le projet européen dans lequel ce travail se situe, maudire l'Université et ses représentants, refuser de continuer, remettre en question tout mon travail, vouloir abandonner et partir loin de tout… certains ne comprenant même plus pourquoi je continuais. Vous avez parfois vu ma santé se dégrader et votre présence m'a, à chaque fois, permis de me remettre sur pieds. Moi-même, je l'admets, j'ai souvent perdu de vue les raisons qui me poussaient à terminer ce travail, accumulant les continuelles frustrations de n'en jamais voir la fin. Malgré cela, j'ai tenu bon, et ce pour deux raisons : la première étant la volonté d'accomplir ce qui a été la tâche la plus ambitieuse de ma vie (et ce ne sera pas la seule, faites-moi confiance), la seconde étant celle, je l'avoue, de vous rendre fier.

Alors à vous, mes Amis, c'est-à-dire *tous ceux qui me sont proches*, sans distinction aucune, je vous remercie une nouvelle fois car chacun, à votre manière — que je n'oublierai jamais —, avez participé à la concrétisation de ce travail.

Merci.

post-scriptum : vous avez intérêt à lire <u>toute</u> cette thèse car je vous interrogerai.

Je tiens également à remercier tous les membres du consortium Two !Ears : Alexander, Hagen et Sandra (Ilmenau et Berlin, Allemagne), Klaus, Ivo, Johannes et Youssef (Berlin, Allemagne), Torsten et Tobias (Copenhague, Danemark), Jens, Dorothea, Thomas et Christopher (Bochum et Berlin, Allemagne), Patrick, Ariel et Thomas (Toulouse, France), Sascha et Fiete (Rostock, Allemagne), Guy et Ning (Sheffield, Grande-Bretagne), Arming et Ryan (Eindhoven, Pays-Bas), Jonaas (Troy, Etats-Unis).

Mais aussi les nombreuses personnes de l'ISIR qui nous aident tous les jours à effectuer notre travail le mieux possible alors que nous, nous ne les aidons jamais... et c'est injuste ! Nommément : Michèle, Anne-Claire, Yves, Ludovic et Adela.

Merci également à Antonyo avec qui j'ai travaillé durant de nombreux mois sur notre cher Odi et qui a rendu possible le « difficilement possible ».

Merci à M. Raja Chatila pour m'avoir accueilli il y a plus de quatre ans et m'avoir permis de venir faire un séminaire à l'ISIR, séminaire qui me permettant, indirectement, de décrocher ce doctorat.

Enfin, Bruno et Sylvain, justement, vous avec qui j'ai travaillé durant ces presque quatre ans, je vous remercie énormément de la confiance que vous m'avez accordée ainsi que de la liberté que vous avez su me laisser, liberté indispensable selon moi à une recherche scientifique pertinente et intègre. J'espère que ce travail de thèse a été à la hauteur de vos exigences et de vos attentes (à vrai dire, j'espère surtout qu'elles les ont dépassées...). J'ai énormément appris à votre contact et je vois désormais le monde un peu différemment, avec des yeux plus perçants et un esprit plus affuté. Je suis également heureux que nous ayons, je pense, dépassé la simple relation doctorant — directeur/encadrant de thèse.

Et aussi, on s'est franchement bien marrés.

post-scriptum : Je remercie aussi les boulangères qui m'ont nourri quasiment tous les jours durant ces quatre ans ainsi que les serveurs et serveuses des différents cafés autour de Jussieu qui m'ont énormément caféiné, tantôt au comptoir tantôt en terrasse. Sans vous, rien de tout ça n'aurait été possible. Rien. Ab-so-lu-ment rien.

« *Nichts ist unergründlicher*

als das System der Motivation unseres Handelns. »

« Rien n'est plus insondable

que le système de motivation régissant nos actions. »

— Georg Christopher Lichtenberg, *Le Miroir de l'Âme*

Table des matières

Chapitre 1

Introduction

Considérons la situation suivante : vous êtes à un dîner entre amis, dans l'appartement de l'un d'entre eux, appartement dans lequel vous n'êtes jamais allé. Vous êtes attablés avec tout le monde dans le salon, discutant, mangeant et buvant. Une des fenêtres est ouverte sur votre droite et vous entendez ainsi les sons provenant de la rue : bruits des voitures et des bus, conversations, chiens qui aboient etc. Soudain, vous entendez une voiture qui freine brusquement et bruyamment. Une de vos premières réactions, si ce n'est la première (tout du moins observable), va très certainement être de *tourner votre tête en direction de la fenêtre afin de* voir *ce qu'il s'est passé*. Nous posons ainsi cette première question :

Quel est le but de cette réaction ?

Ce mouvement de tête a principalement permis de mettre à disposition vos capteurs visuels afin d'effectuer une analyse complémentaire à celle effectuée précédemment par l'audition, modalité sensorielle ayant fourni les informations qui ont provoqué ce mouvement. Tourner sa tête, dans un cas comme celui-ci, est une manifestation du besoin de *porter son attention* sur un événement particulier de l'environnement, étant donné le contexte dans lequel il se situe. Nous posons alors cette deuxième question :

Pourquoi s'intéresser à cette réaction ?

Cette réaction est le résultat de l'intégration de nombreux processus plus ou moins complexes permettant à l'humain de réagir de façon rapide et pertinente dans des environnements complexes. Les mouvements de tête notamment sont déclenchés par un très grand nombre de signaux et de situations : signaux de danger par exemple mais également événements perceptuels inattendus, c'est-à-dire des stimuli requérant notre attention ou porteurs d'un intérêt en regard d'une tâche à effectuer. Dans la situation exposée ci-dessus, au-delà de la notion de signal de danger — entrant également en compte — les caractéristiques principales de cet événement étaient sa rareté, étant donné le contexte, et son *imprédictibilité* relative : aucun indice perceptif, aucune information donnée préalablement n'avait permis d'anticiper l'arrivée de ce son. Mais quelle que soit l'origine des mouvements de tête « réflexe » — excluant

le réflexe moteur visant à dégager la tête d'une situtation de danger physique pour elle — leur point commun est de faire parvenir les capteurs visuels sur une zone de l'espace nécessitant un approfondissement de l'analyse des informations perceptuelles. En effet, la vision étant par nature plus rapide, robuste et précise dans son analyse des signaux lorsque comparée à l'audition, accéder aux informations visuelles permet très souvent de lever des ambiguïtés, améliorer l'analyse de l'environnement ou comprendre la raison de l'imprédictibilité d'un événement. Mais il a également été observé qu'orienter sa tête vers des stimuli sonores permet d'améliorer la perception visuelle.

D'autre part, dans le domaine de la perception sensorielle, une des caractéristiques du cerveau — et sa force — est de pouvoir émetre des *hypothèses* sur ce qui va arriver. Dès lors que c'est possible, et parfois sur la base d'un nombre très faible d'informations, les aires sensorielles vont tenter de prédire les stimuli futurs. La plupart du temps, ces hypothèses sont vérifiées : lancer un verre contre un mur devrait produire un type de son particulier, à une position particulière de l'espace et avec un délai particulier entre le moment où le verre est lancé et celui où il se brise. Cette capacité de prédiction permet (i) d'accélérer grandement les processus d'analyse des informations perçues et (ii) de pouvoir assigner la puissance de calcul de ces aires à d'autres « zones » de l'espace informationnel, moins prédictibles ou plus difficilement analysables. Le cerveau cherche, en d'autres termes, à optimiser le temps de calcul des informations auxquelles il a accès.

Mais lorsqu'un événement imprédictible survient, les aires sensorielles *sur-réagissent*, en réponse à la différence entre la prédiction faite au temps t et l'observation au temps $t+1$. Cette sur-réaction est généralement une commande motrice permettant d'obtenir plus d'informations sur le contexte ayant abouti à cet événement imprédictible, comme, par exemple, générer des mouvements oculaires ou des mouvements de tête. Et au-delà de la simple collecte de nouvelles informations, il va être également question de raffiner le modèle du monde afin d'y inclure ce nouvel exemple afin que cet événement devienne plus *prédictible*. Bien que la rotation de la tête vers un événement perceptif soit en apparence une commande motrice simple, elle permet d'affiner le modèle du monde que le cerveau cherche sans cesse à constuire et à enrichir, de façon rapide et puissante.

Un autre intérêt des mouvements de tête est que ceux-ci possèdent une certaine indépendance par rapport au reste du corps : nous pouvons aller dans une direction tout en en explorant visuellement des parties de l'environnement situées sur les côtés. Ainsi, il est possible d'effectuer deux tâches en parallèle : la navigation vers un point de l'espace et l'exploration d'autres points de l'espace. Or dans le domaine de la robotique humanoïde, nombres d'algorithmes n'incluent pas ce degré de liberté supplémentaire — par contrainte matérielle la plupart du temps — permettant au robot d'augmenter significativement l'aquisition d'informations sans réquisition de la totalité du « corps » du robot.

D'autre part, et c'est le cas qui nous intéresse particulièrement, les mouvements de tête sont également impliqués dans les phénomènes attentionnels (cf. **Fig. 1.1**). Notamment, lorsqu'un stimulus intervient à une position inattendue, un ensemble de neurones organisés sous forme de réseau particulièrement connecté permet de *réorienter* l'attention afin, une fois encore, de comprendre les raisons de l'imprédic-

FIGURE 1.1 – MOUVEMENTS DE TÊTE ET ATTENTION — Phénomène de *réorientation de l'attention* dans lequel un mouvement de tête permet de réquisitionner les capteurs visuels afin qu'ils puissent acquérir des données issues d'un événement d'intérêt. Ce phénomène a des bases neurales dans diverses aires cérébrales, notamment les aires frontopariétales, comme expliqué à la **Sec. 2.3** (figure d'après [1]).

tibilité de cet événement.

Toutes ces considérations ont servies de motivation au développement du modèle HEAD TURNING MODULATION conférant à un robot mobile la capacité de générer de façon indépendante des mouvements de tête. Le travail présenté ici a de plus fait partie du projet FET [1] européen TWO!EARS, ayant démarré le 1er décembre 2013 et s'étant terminé le 30 novembre 2016. L'ambition des dix laboratoires composant le consortium TWO!EARS a été d'intégrer au sein d'une plateforme robotique mobile dotée d'une audition binaurale et d'une vision binoculaire une architecture logicielle puissante incluant, notamment, des modèles innovants d'audition binaurale active. Or la binauralité, dans le domaine de la robotique, est une approche plutôt peu commune. En effet, même si les performances de l'humain sont très bonnes (cf. notamment dans le fameux cas du *cocktail party*, décrit plus tard), les systèmes robotiques binauraux actuels sont globalement peu performants lorsqu'il est question d'analyser des scènes au contenu acoustique complexe, particulièrement en terme de nombres de sources sonores simultanées ou de conditions réverbérantes. Mais alors pourquoi ne pas simplement doter le robot d'une dizaine de microphones ? Le choix de l'audition binaurale est motivé par de multiples et diverses raisons, une des principales pouvant être formulée par la question suivante : cherche-t-on à doter le robot d'une audition parfaite ? Ou cherche-t-on plutôt à comprendre comment l'humain peut réaliser de si bonnes performances avec seulement deux oreilles, compréhension qui sera ensuite utilisée par la communauté robotique, entre autres, pour créer des modèles bio-inspirés (modèles qui, par ailleurs, ont l'avantage d'avoir également un impact sur la communauté biologique : nombre de modèles computationnels de mécanismes cérébraux ont eu pour conséquence l'influence de travaux de recherche de neurosciences permettant de mettre au jour de nouvelles structures cérébrales ou neuronales) ?

1. Future and Emergent Technologies

Mais ces performances ne sont pas dues qu'à notre très bonne capacité à analyser des signaux audio, ou à nos capteurs — nos oreilles — dotés d'une sensibilité particulière. Selon nous, tenter de résoudre des situations acoustiques très complexes n'est seulement possible qu'avec le concours de la vision. D'ailleurs, toujours dans le problème du *cocktail party*, cette modalité joue un rôle prépondérant dans la discrimination de sources sonores ou dans l'amélioration de la traque de signaux audio permettant une facilitation de leur identification et de leur localisation. En effet, la vision permet de localiser une source audiovisuelle par exemple, information que le système auditif peut utiliser pour affiner son analyse : *voir* une source émettre un son face à nous permet de simplifier grandement l'analyse de la scène acoustique en supprimant toutes les informations parasites gênant éventuellement le traitement des signaux acoustiques bruts (réflections multiples, réverbération, cône de confusion, ambiguïté avant-arrière etc.). Nous revenons ainsi ici aux mouvements de tête, mouvements qui permettront d'inclure la vision et d'acquérir des informations supplémentaires issues d'une modalité différente afin d'améliorer l'analyse de scène auditives. Mais un des atouts majeurs de l'audition est sa capacité à capter des sons dans un champ de 360°. Les oreilles peuvent en effet percevoir des stimuli tout autour de la tête (tant qu'il n'y a pas d'obstacles à la propagation du son), là où la vision n'est que limitée à un champ restreint. Tirant partie de cette différence, mais gardant toujours en tête que la vision est un support d'analyse plus précis que l'audition, particulièrement dans le monde robotique, nous avons considéré la modalité auditive comme un *déclencheur de réactions motrices* : les mouvements de tête. Ainsi, à chaque fois qu'un stimulus audio présente un intérêt pour le robot — et nous définirons précisément la façon dont nous avons formalisé cette notion d'intérêt au sein de notre modèle — un mouvement de tête sera généré vers l'objet en question.

Le modèle Head Turning Modulation (HTM) a donc pour ambition de doter un robot d'une capacité à tourner sa tête lorsque des événements d'intérêt surviennent dans un environnement. Pour cela, il tente de permettre à ce robot de se construire de façon autonome et non-supervisée une représentation interne des environnements qu'il explore sur la base des objets multimodaux qui y sont présents. Afin de parvenir à ce but, le modèle HTM va analyser l'apparition d'événements audiovisuels selon leur *Congruence* à l'environnement dans lesquels ils se situent. Cette notion de Congruence sera un des versants du concept d'intérêt d'un événement pour le robot et sera décrite en détail dans un chapitre dédié à la partie du modèle en charge de cette analyse.

Pour résumer, la question à laquelle tente de répondre le modèle HTM est ainsi la suivante :

> *Comment faire en sorte qu'un robot mobile doté d'une perception audiovisuelle similaire à l'Homme puisse de lui-même comprendre ce qu'est un* stimulus d'intérêt *afin de porter son attention dessus, tout en restant sensible à tous les autres stimuli présents dans un environnement inconnu en cours d'exploration ?*

Cette question fait appel à un nombre important de notions et de concepts tant en rapport avec la robotique, l'ingénierie, qu'avec les sciences du comportement ou

la neurologie. Parmi ceux-ci : exploration d'un environnement inconnu, perception multimodale et active, stimulus d'intérêt, attention et cognition. Le modèle HTM est ainsi au carrefour de nombreux domaines de recherche, tous vastes et complexes. Mais à la question posée plus haut, le présent travail, et plus particulièrement le modèle qui en est sa concrétisation, tentera d'y répondre en considérant que des comportements complexes, faisant tout aussi bien appel à l'intégration multimodale, la perception active, l'apprentissage en temps réel, voire même, des prémisses de conscience [2, 3], peuvent être modélisés par des systèmes simples mais dont les bases conceptuelles sous-jacentes sont suffisamment solides pour rendre compte de ce qu'on peut appeler une forme d'intelligence robotique.

Ce manuscrit de thèse est organisé selon les chapitres suivants :

Etat de l'art : ce chapitre synthétise l'ensemble des concepts et des notions sur lesquels le modèle HTM se base (Exploration, Perception et Attention, Apprentissage et Fusion de classifieurs). Les domaines auxquels le modèle HTM est lié sont nombreux. Ainsi, seulement les travaux de recherche les plus importants et les plus pertinents seront détaillés ici.

Two !Ears : le modèle HTM a fait partie du projet européen TWO!EARS. Ce chapitre décrira ainsi en détail le projet, son architecture, les deux plateformes robotiques utilisées ainsi que les scénarios de tests créés pour valider l'ensemble du logiciel TWO!EARS.

Introduction au modèle HTM : ce chapitre introduira le modèle HEAD TURNING MODULATION en posant des définitions et des notations primordiales pour comprendre le fonctionnement des deux modules constitutifs du modèle. De plus, la notion d'*Objet* et la façon dont elle a été comprise au sein de notre modèle sera détaillée. Enfin, l'environnement de simulation créé pour tester le modèle HTM, le HtmTestBed, sera décrit ainsi que les critères d'évaluation du modèle que nous avons utilisés.

Module de Pondération Dynamique : il s'agit du module de base du modèle HTM. Il est en charge de l'analyse des événements audiovisuels en fonction de leur Congruence, notion qui sera également définie et formalisée dans ce chapitre. L'évaluation du module sera également effectuée.

Module de Fusion et d'Inférence Multimodale : ce module a été implémenté afin de résoudre le problème d'inférence de données manquantes et de fusion multimodale : comment retrouver, par exemple, l'information visuelle d'un objet situé derrière le robot ? L'évaluation du module sera également effectuée.

Combinaison des modules et Intégration sur le robot : la combinaison des deux modules, rendant le modèle complet sera ici décrite. Son adaptation en tant qu'expert computationnel afin de l'intégrer au système TWO!EARS porté sur le robot sera présentée. Les évaluations du modèle entier, en simulation, puis de sa version TWO!EARS (en tant que KS) intégrée au vrai robot seront effectuées ici.

Conclusion : la conclusion sur ces trois années de thèse permettra de résumer l'ensemble du travail qui a été effectué mais aussi les limites du modèle ainsi que les différentes améliorations qui peuvent y être apportées.

Remarques préalables à la lecture de ce manuscrit :

- Ce document décrit la majeure partie du travail qui a été effectué durant cette thèse. Bien que long, il constitue malgré tout une vraie synthèse de tout ce qui a été accompli. Nous souhaitons avertir le lecteur sur le fait que, bien que la description du cœur du modèle HTM arrive relativement tard dans ce manuscrit, nous avons jugé indispensable de présenter préalablement tous les concepts, notions, définitions et cadres ayant motivés, influencés ou aidés le travail de thèse présenté ici.
- L'état de l'art est dense car il couvre un ensemble de domaines différents, chacun étant séparément très vaste. Cet état de l'art est également une synthèse visant à présenter toutes les sources d'inspirations de ce modèle. Afin d'aider le lecteur, nous l'avertissons que les parties 1 et 3 de l'état de l'art concernent principalement le module DW (**Chap. 5**) et que les parties 2 et 4 concernent plutôt le module MFI (**Chap. 6**). Ainsi, il sera tout à fait possible de lire ce chapitre d'état de l'art de façon épisodique et non de façon linéaire.
- Nous avons décidé de présenter le projet Two!Ears après le chapitre d'état de l'art. Bien que cette thèse soit incluse dans ce projet européen et donc soumise à de nombreuses contraintes autres que celles que nous avons posées de notre côté, ce travail de thèse constitue néanmoins un travail de recherche avant tout. C'est pourquoi nous avons souhaité mettre d'abord l'accent sur les nombreuses travaux ayant inspirés ou concurru à la création et au développement du modèle HTM *via* l'état de l'art, puis de présenter le projet Two!Ears.
- Enfin, le modèle HTM étant constitué de deux modules différents mais dont certaines bases formelles sont partagées, nous avons choisi de commencer la description du modèle par un chapitre introductif (**Chap. 4**) sur lequel les deux chapitres suivants, dédiés aux deux modules justement, se basent. Lors de la lecture du **Chap. 5** et du **Chap. 6**, le lecteur est ainsi invité à se référer au **Chap. 4** pour les définitions et notations sur lesquelles la suite du document se base.

Positionnement du Problème

Question posée

Comment faire en sorte qu'un robot mobile doté d'une perception audiovisuelle similaire à l'Homme puisse de lui-même comprendre ce qu'est un *stimulus d'intérêt* afin de porter son attention dessus, tout en restant sensible à tous les autres stimuli présents dans un environnement inconnu en cours d'exploration ?

Description concrète du problème

Un robot doté de mouvements de tête et de capteurs audio et visuels explore des environnements inconnus dans lesquels sont placées des sources audiovisuelles. A partir des données issues de divers « experts », notamment ceux dédiés à l'identification et la localisation de ces sources audiovisuelles et fournis par le système Two!Ears, le modèle proposé ici va tenter de générer des mouvements de tête vers certaines sources. La détermination de ces sources d'*intérêt* se fait grâce à deux types d'analyses :

- une analyse sur la base du contenu sémantique des données, à savoir leur catégorie audiovisuelle, permettant de déterminer si la source présente un *intérêt* ou non pour le robot (module Dynamic Weighting, DW),
- une analyse de la qualité de la connaissance que le système a de l'environnement (module Multimodal Fusion & Inference, MFI), notamment sa capacité à (i) inférer une information manquante (audio ou visuelle) et (ii) corriger les éventuelles erreurs de classification, assurant ainsi au DW de toujours avoir accès à une représentation audiovisuelle pertinente des objets perçus.

Chacune de ces analyses va pouvoir, séparément, déterminer si une des sources présentes dans l'environnement nécessite un mouvement de tête.

Plateforme robotique utilisée

- robot mobile,
- audition binaurale,
- vision binoculaire,
- cou permettant des rotations de la tête en azimut.

Champs de recherche impliqués

- exploration,
- perception,
- attention,
- apprentissage,
- fusion de données.

Chapitre 2

Etat de l'Art

Ce chapitre regroupe toute la littérature qui a servie de base à la conception du modèle Head Turning Modulation. Les inspirations de ce modèle sont nombreuses et variées, passant des phénomènes attentionnels à la robotique mobile. Durant tout ce travail de thèse, une importance majeure a été donnée à la compréhension et la connaissance des phénomènes cérébraux responsables de l'exploration, de la perception ou de l'attention. C'est pourquoi les différentes sections de ce chapitre d'état de l'art seront systématiquement ouvertes par une description des travaux de recherche consacrés aux aspects neuronaux et/ou cognitifs servant ensuite de base à la conception de modèles computationnels ou d'implémentations robotiques.

Avant de procéder à cet état de l'art, et étant donnée l'étendue des domaines dans lequel le modèle HTM est impliqué, il est nécessaire de définir le modèle dans ses grandes lignes. Le modèle HTM est un modèle de *modulation des mouvements de tête d'un robot dans le cadre de l'exploration d'un environnement inconnu*. Cette modulation des mouvements de tête sera effectuée en réponse à des *Motivations* telles que la réduction de l'incertitude sur l'environnement ou la congruence des événements apparaissant dans cet environnement. D'autre part, dans la formalisation HTM, l'environnement est défini par les objets audiovisuels qui le composent. Ainsi, le modèle comporte une partie d'intégration multimodale des données sensorielles perçues par le robot, ainsi que, couplée à cette partie, la capacité d'inférer des données issues d'une modalité manquante. Un des buts du modèle est de conférer au robot une représentation interne des environnements qu'il aura exploré, de façon toujours centrée sur la notion d'objet audiovisuel. De la volonté de créer cette représentation sous forme de carte cognitive, le modèle intègrera un algorithme d'apprentissage des données servant autant à l'apprentissage des données multimodales qu'à une forme de mémoire à long-terme. Plusieurs domaines sont donc concernés par le modèle HTM : exploration, perceptions auditive et visuelle, phénomènes attentionnels, intégration multimodale, apprentissage etc.

En premier lieu, la **Sec. 2.1.1** consistera en une description des processus cérébraux responsables de la navigation, la localisation dans un environnement, ainsi que son exploration. Ensuite, la **Sec. 2.1.2** décrira les principales méthodes d'exploration utilisées en robotique mobile, et particulièrement celles inspirés du vivant. La section **Sec. 2.1.3** sera quant à elle dédiée aux principaux travaux de recherche portant sur

la formalisation et l'implémentation de systèmes intégrant des formes de *Motivation* à l'exploration.

D'autre part, le modèle HTM se base sur une perception audio et visuelle de l'environnement. La **Sec. 2.2** décrira ainsi la perception audio, visuelle ainsi que leur intégration multimodale. De plus, certaines caractéristiques cognitives de ces phénomènes, notamment la Théorie de la Hiérarchie Inverse seront introduites à la **Sec. 2.2.3**. D'autre part, le modèle étant voué à fonctionner sur un robot doté de mouvements de tête, la **Sec. 2.2.1.3** introduira les travaux menés sur l'apport de ce type de mouvements pour la perception, auditive particulièrement.

La faculté de tourner sa tête en fonction de l'apparition de certains stimuli audio ou visuels fait appel à la notion d'*Attention*. La **Sec. 2.3** détaillera ainsi les structures cérébrales impliquées particulièrement dans l'attention « ouverte » (i.e. visible) incluant donc la génération de commandes motrices vers des stimluli d'intérêt. De plus, la notion de *Saillance* d'un stimulus, sur laquelle la notion de Congruence se base, sera introduite à la **Sec. 2.3.1.2**.

Enfin, le modèle HTM inclut un algorithme d'apprentissage des données perçues. La **Sec. 2.4** détaillera ainsi les principaux paradigmes employés dans ce domaine : apprentissage supervisé, non-supervisé et auto-supervisé, avec une emphase sur le dernier type puisque c'est celui qui a été utilisé dans le cadre de l'élaboration du modèle Head Turning Modulation. La **Sec. 2.4.1.1** détaillera quant à elle le fonctionnement des cartes auto-adaptatrices, réseaux de neurones ayant servi de base à partir de laquelle un algorithme d'apprentissage plus complexe a été développé. Enfin, une étape importante du processus d'analyse des signaux audiovisuels perçus par le robot étant la fusion des modalités audio et visuelle, la **Sec. 2.4.2** introduira les stratégies de fusion de classifieurs.

2.1 Exploration

L'exploration d'un environnement inconnu peut être définie, d'après les travaux de John O'Keefe & Lynn Nadel [4] sur le comportement exploratoire des rats, comme :

> *La suite d'actions motrices visant à acquérir toute information pertinente et robuste permettant d'obtenir une représentation sous forme de carte et qui pourra être utilisée ultérieurement afin de s'y localiser ou d'y planifier un déplacement.*

Cette exploration peut aussi bien porter sur l'espace (topologie du terrain, obstacles, chemins empruntables etc.) que sur le contenu (présence d'objets, notion de danger etc). Par exemple, percevoir, dans un environnement inconnu, un homme qui parle situé à notre gauche et placé derrière un bureau, constitue aussi bien une exploration de l'espace (position de l'homme et obstacle à contourner pour l'atteindre) que du contenu sémantique (action/comportement de l'homme). Si le but de l'exploration est d'être capable de parcourir l'environnement afin de le cartographier, il vient qu'une des étapes primordiales et essentielles de l'exploration est donc la bonne

connaissance de cet environnement, de sa topologie. En addition de cette exploration topologique, les informations perçues sur les *entités* [5] statiques ou non-statiques, telles que l'homme qui parle, constitue une exploration sémantique et peut/doit être utilisée pour prendre des décisions pertinentes permettant d'adopter une stratégie réactive efficace. En reprenant l'exemple ci-dessus, ce type d'événement pourrait représenter un intérêt et devenir une motivation pour aller dans la direction de l'homme. D'autre part, si, sur ce chemin tout juste planifié, est détecté un enfant qui pleure à l'opposé de l'homme, peut-être serait-il pertinent de prioriser ce nouvel événement et ainsi mettre à jour la prochaine position à atteindre. Cet aspect dynamique de l'exploration englobe donc les concepts traditionnels de l'exploration d'un environnement, tels que la *navigation*, la *localisation* et la *cartographie*, mais également des concepts plus récents et faisant appel à un aspect plus cognitif de l'exploration, tels que la *motivation* et la *prise de décision*. La prochaine section constitue donc un état de l'art sur ces différents volets de l'exploration : exploration spatiale chez l'animal et bases neurales sous-jacentes, stratégies d'exploration robotique et modèles bioinspirés, et motivations pour l'exploration.

2.1.1 Bases neurales de l'exploration, de la navigation et la localisation

> *Q. : Quelles sont les bases neurales responsables de la capacité qu'ont les animaux à explorer des environnements inconnus de façon rapide, robuste et pertinente ?*

L'exploration d'un environnement nécessite tout d'abord d'avoir la capacité de se localiser dans cet environnement, afin, dans un second temps, de pouvoir s'y déplacer. Deux stratégies principales sont utilisées par les animaux pour naviguer au sein d'un environnement :

- *navigation par repères* ou *allothétique* : l'animal infère sa position et son orientation grâce à la détection de repères purement externes par la vision, l'audition, l'olfaction etc. ;
- *intégration spatiale* ou *navigation idiothétique* : l'animal connaît sa position et son orientation de départ et les estime ensuite sur la base de repères internes comme la kinesthésie, la copie efférente ou les informations vestibulaires et proprioceptives.

Ces deux types de stratégies diffèrent par l'information préalable que l'animal a sur sa propre position dans l'environnement : dans le cas allothétique, les informations spatiales utilisées lui sont extérieures, tandis que dans le cas idiothétique, les informations spatiales sont totalement internes. Conséquemment à ces deux stratégies globales de navigation, deux représentations différentes de l'environnement sont créées lors de cette navigation :

- *allocentrique* : représentation indépendante de la position de l'animal,
- *égocentrique* : représentation dans laquelle les indices spatiaux perçus sont relatifs à l'animal.

Ces deux représentations peuvent également être comprises selon leur caractère absolu (allocentrique) ou relatif (égocentrique). Les informations spatiales, quelles

soient allothétiques ou idiothétiques, ne sont pas simplement des données aquises au cours de la navigation sur la position d'objets ou la position de l'animal à un temps donné. Ces informations sont stockées et intégrées au cours du temps afin de créer une représentation stable et réutilisable, généralement formalisée sous forme de *carte*.

En 1978, John O'Keefe & Lynn Nadel [4] ont émis l'hypothèse qu'une carte cognitive devait exister dans le cerveau du rat (animal intensivement étudié pour ses excellentes capacités d'exploration et de navigation) lui permettant de se localiser et d'explorer des environnements inconnus de façon rapide et performante. L'hippocampe (HS), en tant que structure prépondérante impliquée dans la mémoire ainsi que dans de nombreux autres processus est fortement mobilisée lors des processus de localisation et de navigation en cela qu'il permet d'avoir accès à des informations spatiales précédemment collectées, analysées et organisées. HS est situé dans le lobe temporal médian, sous la surface du cortex et est présente dans les deux hémisphères (structure cérébrale paire). Composé du gyrus denté, du subiculum et de la corne d'Ammon, HS interagit notamment avec le cortex entorhinal (EC), le cortex préfrontal (PFC) et le noyau accumbens (NA). Ces aires sont les principales structures cérébrales responsables de la capacité d'un individu à s'orienter, au sens large, dans un environnement (voir [6, 7, 8, 9, 10, 11] et [12] pour une revue). O'Keefe & Nadel, ont mis en évidence le fait que l'hippocampe joue un rôle majeur dans l'élaboration d'une carte cognitive spatiale, notamment *via* l'existence de *Place cells* (PC), mises en évidence en 1971, par John O'Keefe *et al.* [13]. Les PC s'activent lorsque l'animal est dans une zone particulière de l'environnement et sont stables dans le temps [14]. Dans des environnements inconnus, ces cellules sont rapidement recrutées afin de coder au plus vite l'information spatiale acquise lors de la navigation. De plus, les PC ne sont pas liées à une modalité précise : Robert U. Muller & John L. Kubie, en 1987 [15], Gregory J. Quirk *et al.* en 1990 [16] ou encore Ethan J. Markus *et al.* en 1994 [17] ont montré qu'elles étaient actives dans des environnements sombres, chez des individus aveugles, ou pouvaient même être activées par l'odorat. De façon étonnante, une des particularités des PC est qu'elles peuvent également être activées dans deux environnements différents [18]. Ainsi, la notion de position n'est pas directement liée à l'environnement mais plutôt à sa perception puis à sa représentation. D'autre part, les PC ne sont pas exclusivement excitées par des stimuli spatiaux : Robert E. Hampson *et al.* [19] en 1993, a montré que ces cellules peuvent s'activer en réponse à des événements saillants au sein d'une séquence temporelle ; Brian G. Young *et al.* [20] en 1994 et Emma R. Wood *et al.* [21] en 1999 ont montré que des stimuli comme la texture ou l'odeur peuvent aussi déclencher une activation des PC.

Cependant, en 2001, David A. Redish [22] met en avant les limites de l'approche de John O'Keefe : l'hippocampe seul ne suffit pas à créer une carte cognitive allothétique de l'environnement. Il serait plutôt impliqué dans la combinaison de plusieurs cartes, définies comme des associations entre des indices spatiaux externes et un système de coordonnées intrinsèque à l'animal, système mis à jour par les informations vestibulaires, la proprioception et les structures impliquées dans le phénomène de copie efférente. L'hippocampe ne serait qu'une partie d'un système vaste et complexe impliquant de nombreuses aires corticales et structures cérébrales différentes et prenant en compte un grand nombre d'informations comme le mouvement de la

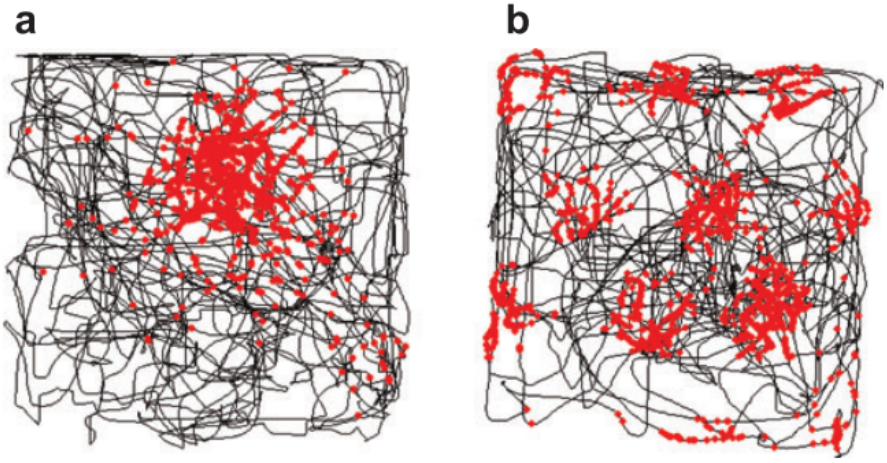


FIGURE 2.1 – GRID CELLS — (*a*) Cellules de type *Place cells* dans l'hippocampe ; (*b*) cellules de type *Grid cells* situées dans le cortex entorhinal. La localisation des pics neuronaux (*spikes*), dénotés en rouge, se superposent à la trajectoire de l'animal dans l'enceinte (*traits noirs*). Alors que la plupart des cellules ne s'activent qu'en un point de l'environnement, les *Grid cells* s'activent selon une matrice triangulaire périodique couvrant l'espace exploré (figure d'après [24])

tête, les informations motrices, la perception visuelle et la tâche à accomplir. Réduire les capacités d'exploration d'un environnement au seul hippocampe n'est donc pas suffisant pour les expliquer. Cependant, les *Place cells* restent un des grands groupes de cellules sur lesquels la localisation et la navigation se basent. En effet, le codage de la position spatiale, même s'il est soumis à de très nombreuses afférences neuronales, est le cœur du système d'exploration animal.

En 2005, TORKEL HAFTING *et al.* [23] mettent en avant les *Grid Cells* (GC), cellules situées dans EC, juste sous HS. Les GC fonctionnent comme des champs récepteurs en cela qu'elles forment des connexions triangulaires, appelées *grilles*, codant des zones de l'environnement que l'animal a déjà exploré (cf. **Fig. 2.1**). Cependant, et contrairement à la plupart des cellules neuronales, les GC s'activent par groupe formant une matrice triangulaire couvrant un espace plus grand que celui occupé par l'animal. Cette organisation particulière est ainsi supposée être une forme de système métrique utilisée dans l'analyse des informations spatiales et particulièrement lors de la navigation. En effet, les GC sont caractérisées par :

- *l'espace*, exprimé comme la distance entre les champs récepteurs,
- *l'orientation*, définie comme l'inclinaison relative en fonction d'un axe de référence externe,
- *la phase*, définie comme le déplacement en xy également relatif à un point de référence externe.

L'hippocampe et le cortex entorhinal sont deux structures majeures dans l'élaboration d'une représentation interne d'un environnement, par le moyen de cartes spatiales, ou cartes cognitives. Au sein de ces deux structures complexes, les PC et les GC sont parmi les éléments constitutifs de ces cartes. Il est également intéressant de citer les *Head-Direction cells*, présentes dans le présubiculum [25, 26] et dans le thalamus antérieur [27], et modulant l'activité des GC en fonction de l'orientation

de la tête.

D'autre part, le principe de *navigation* englobe autant la possibilité de se repérer dans un environnement que de planifier un trajet vers une zone cible. Cette planification est complexe car motivée par de nombreux processus qui vont de l'action réflexe au comportement complexe et intégré, multimodal et fondamentalement cognitif. Cependant, la planification d'une stratégie de navigation fait également appel à la détermination d'une suite d'actions à effectuer afin d'aller d'un point à un autre de l'environnement. En amont de l'aspect décisionnel haut-niveau consistant à choisir quel chemin pourrait être le plus « bénéfique » pour l'animal, selon le but global de l'exploration, peut exister un conflit plus bas-niveau entre l'accomplissement de deux actions antinomiques, comme celle d'aller à gauche ou à droite. Etienne Koechlin & Yves Burnod [28] définissent une *action* comme l'activité globale de réseaux neuronaux la représentant, activité neuronale présente dans tout le système nerveux central. Autrement dit, un ensemble de neurones appartenant à différentes aires cérébrales impliquées dans divers aspects de l'analyse des informations perçues et dans l'éventuelle réaction consécutive requise, va globalement représenter une *action*. Une action comme celle de « tourner à gauche » du fait de la perception d'un stimulus audio d'intérêt fera aussi bien appel aux neurones impliqués dans l'exécution d'une action motrice, que ceux impliqués dans l'analyse du son et de son « intérêt » (notion de récompense par exemple) à, que ceux également impliqués dans l'anticipation des informations qui seront perçues à la suite de ce mouvement. Cet ensemble peut être ainsi compris comme une *action*.

La sélection de la prochaine action a exécuter peut alors être déterminée en analysant l'activité des réseaux de neurones, appelés parfois canaux d'information, représentant les actions motrices candidates à une sélection en vue de leur exécution [29]. Cette activité est d'ailleurs définie comme une *saillance* ou comme la *propension d'une action à être exécutée* [28]. Chez l'homme, c'est dans les ganglions de la base (GdB) notamment que se trouve le mécanisme responsable de la *sélection de l'action* [30, 31, 32], structure composée de plusieurs sous-structures (le striatum, le pallidum et le noyau subthalamique — chacune d'entre elles se décomposant à nouveau en d'autres sous-structures). En 1999, Tony J. Prescott *et al.* [33] et Peter Redgrave *et al.* [34] ont proposé une hypothèse unifiée du phénomène de sélection de l'action par les ganglions de la base par regroupement d'études physiologiques et anatomiques : selon leur hypothèse, chaque action arrive dans les ganglions de la base en étant inhibée par défaut et les GdB vont lever l'inhibition de l'action ayant la *saillance* la plus forte déclenchant ainsi l'ordre moteur associé.

Discussion

De nombreuses structures cérébrales jouent un rôle prépondérant dans la localisation et la navigation, comme illustré à la **Fig. 2.2**. Parmi celles-ci, l'hippocampe, le cortex entorhinal et les ganglions de la base sont celles possédant les réseaux neuronaux et les types cellulaires les plus importants. De plus, les ganglions de la base permettent de mettre en relation différentes afférences motrices éventuellement contradictoires et participe à la sélection des actions motrices possibles, filtrage au cœur de la notion d'exploration d'un environnement inconnu. La complexité de toutes les connexions existantes entre ces différentes structures cérébrales, couplées

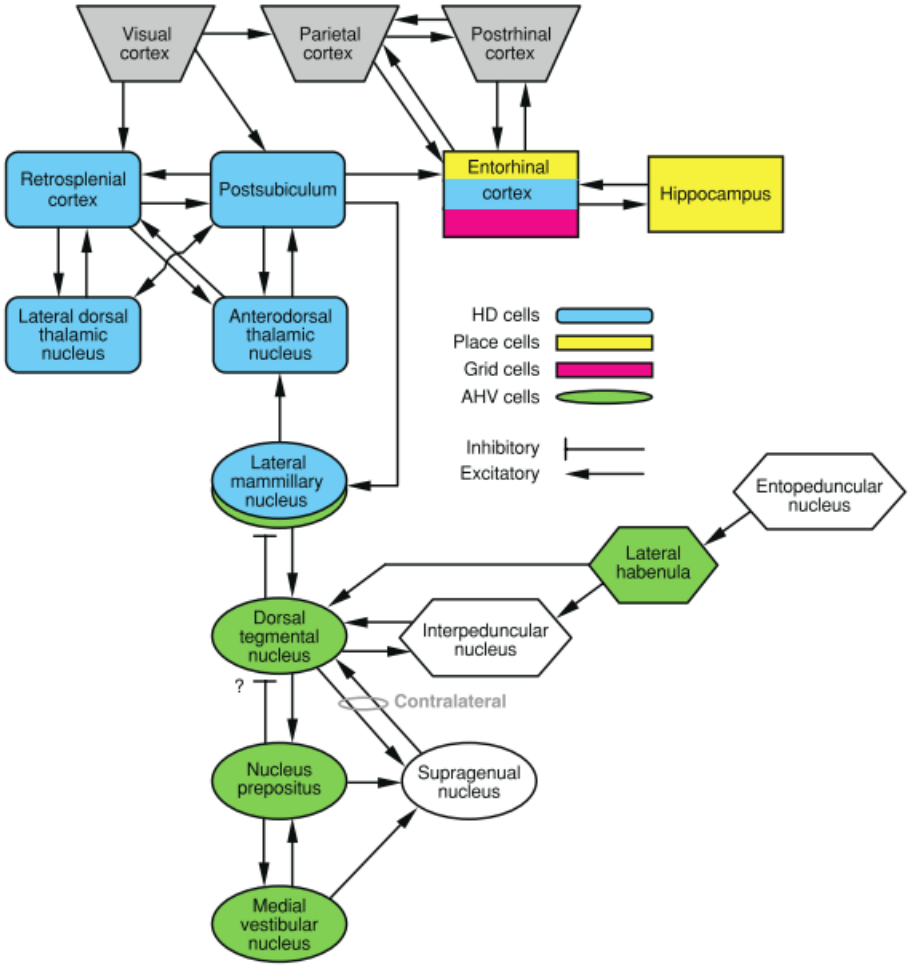


FIGURE 2.2 – RÉSUMÉ DES STRUCTURES CÉRÉBRALES IMPLIQUÉES DANS L'EXPLORATION — Diagramme présentant les connexions principales entre les aires cérébrales contenant les différents types de cellules impliquées dans l'exploration, la navigation et la localisation chez l'animal (figure d'après [35]).

à des neurones au comportement fortement spécialisé et activés dans des conditions bien définies, a entraîné la communauté robotique a se concentrer premièrement dans l'implémentation de systèmes inspirés de l'hippocampe et du cortex entorhinal, en particulier la modélisation computationnelle des PC et des GC. D'autre part, des modèles fonctionnels des ganglions de la base permettent de doter les algorithmes d'exploration d'une architecture neuronale computationnelle robuste et adaptative aux différents buts globaux d'une exploration d'un environnement inconnu.

Nous avons également vu, en introduction à ce chapitre, que la navigation animale peut être séparée en deux grandes méthodes : la navigation allothétique et la navigation idiothétique. A ces méthodes sont liées les représentations allocentrique ou égocentrique de l'animal au sein de la représentation qu'il se fait de l'environnement et des informations qu'il utilise à cette fin. Le modèle HTM a accès aux données odométriques (équivalent robotique de la proprioception) ainsi qu'aux données externes (position spatiale des objets détectés par les capteurs du robot). Nous avons opté pour une représentation égocentrique de l'environnement : le robot est au centre du repère spatial et toutes les informations qu'il perçoit sont ainsi exprimées selon son

propre référentiel.

Par ailleurs, nous avons déjà constaté déjà les premières occurrences de la notion de *Saillance*. Cette notion sera présente dans quasiment toutes les sections de cet état de l'art bien que concernant des domaines très différents. Il s'agit d'une des bases conceptuelles les plus importantes du modèle HTM : cette notion de saillance, adaptée à notre problème, va être une des motivations pour l'exploration d'une zone particulière de l'environnement et ce, de façon comparable à la façon dont les ganglions de la base traitent le problème de sélection de l'action.

Mais avant de détailler cette partie du modèle, la section suivante va se pencher sur les principaux travaux de recherche en robotique mobile s'inspirant des structures neuronales et/ou cérébrales précédemment détaillées : *Place cells*, *Grid cells*, hippocampe ou ganglions de la base.

2.1.2 Application en Robotique

Selon Alexei A. Makarenko *et al.* [36], l'exploration d'environnements inconnus par un robot implique d'être capable d'effectuer principalement trois tâches (cf. **Fig. 2.3**) :

la cartographie qui est l'intégration des informations perçues par les capteurs du robot dans une représentation,
la localisation qui est l'estimation de la postion du robot,
le contrôle moteur qui pose le problème de la façon dont le robot doit effectuer une suite d'actions motrices afin de parvenir à une position déterminée.

De nombreuses techniques et algorithmes ont été créés dans le but d'implémenter des stratégies d'exploration efficaces. Il est également possible de les classer selon deux catégories : exploration rapide et quantité maximale d'information. Les stratégies d'exploration rapide sont un ensemble de techniques visant à minimiser le temps nécessaire pour une exploration de l'environnement entier, incluant donc également une minimisation du temps consacré au calcul des ordres moteurs optimaux permettant d'atteindre ce but. Les stratégies basées sur la quantité maximale d'information sont un ensemble de techniques visant à réduire au maximum l'*incertitude* à propos de l'environnement exploré en choisissant le prochain point d'observation qui maximisera l'acquisition d'information nouvelle. L'accent est donc ici mis sur l'exploration, la contrainte temporelle devenant ainsi secondaire. Cette exploration sera, pour la plupart des modèles computationnels, basée principalement sur des *repères visuels* situés dans l'environnement comme des objets, des personnes ou des parties du décor. Ils seront utilisés par le robot pour créer une carte interne lui permettant de s'y situer et d'y naviguer. Par exemple, de tels repères sont utilisés pour résoudre le problème dit de *loop-closing*[1] qui est la capacité — ou l'incapacité — d'un robot à reconnaître une zone de l'environnement déjà explorée [37].

Afin de développer des algorithmes d'exploration performants, une partie de la communauté robotique s'est souvent inspirée du monde animal. L'exploration d'un environnement inconnu est une tâche que la plupart des animaux effectuent et dans

1. *Fermeture de boucle*

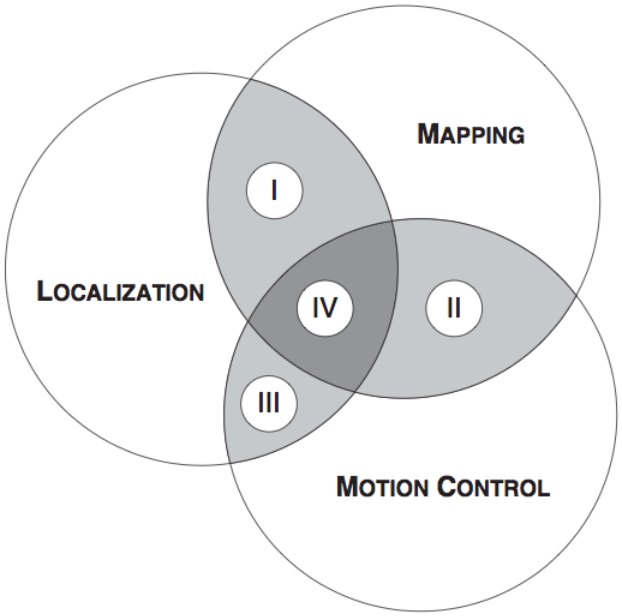


FIGURE 2.3 – TÂCHES IMPLIQUÉES LORS DE L'EXPLORATION ROBOTIQUE — Illustration des trois tâches qu'un robot doit accomplir afin d'explorer un environnement inconnu de façon pertinente et efficace : localisation, cartographie et contrôle moteur. Les intersections correspondent à l'intégration de ces tâches en robotique : (I) Localisation et cartographie simultanée (SLAM, voir **Sec. 2.1.2.1**) ; (II) exploration classique ; (III) localisation active ; (IV) exploration intégrant ces trois tâches (figure d'après [36]).

laquelle ils montrent des performances ainsi qu'une facilité, apparente, étonnantes. De plus, ils sont capable de s'adapter très facilement à des nouveaux environnements. Ainsi est née, au tout début des années 1990, et notamment grâce à JEAN-ARCADI MEYER & AGNÈS GUILLOT [38] l'approche *Animat*, motivé par la volonté d'implémenter des systèmes robotiques (i) imitant le comportement des animaux, en particulier lorsque ceux-ci sont placés dans des environnements imprédictibles et potentiellement dangereux, et (ii) dotés de capacités d'apprentissage. Les *Animats*, contraction d'***ani****mal-****mat****erials*, sont des animaux artificiels, qu'ils soient simulés ou physiques. Le comportement des animaux est un challenge pour la communauté robotique qui, lorsqu'il est question d'exploration d'environnements par des agents robotiques, se basent principalement sur deux approches :

- un apprentissage *supervisé* entrainant de bonnes performances dans les environnements appris mais de mauvaises performances dans des environnements inconnus ;
- un apprentissage *non-supervisé* mais imposant de très fortes contraintes ainsi qu'un ensemble de définitions putatives de l'environnement à explorer, ensemble donné *a priori*.

La **Sec. 2.1.2.1** introduira un des principaux paradigmes utilisé en exploration robotique et employée au sein de TWO!EARS, la technique SLAM ; puis la **Sec. 2.1.2.2** détaillera des modèles bio-inspirés d'exploration d'environnements, notamment l'approche Animat (RatSlam et PsiKharpax) ainsi que le modèle GPR.

2.1.2.1 Localisation et cartographie simultanée — SLAM

Le problème dit de « localisation et de cartographie simultanée » peut être formulé par la question suivante [39] :

> *Est-il possible qu'un robot mobile soit placé à une position et dans un environnement inconnus et qu'il puisse construire par lui-même, de façon incrémentale, une carte de cet environnement, tout en déterminant sa localisation ?*

Cette question est intéressante en cela que pour qu'un robot puisse se localiser, il a besoin d'une carte, et que pour construire une carte, il a besoin de savoir à partir de quels point de l'environnement ont été perçus les répères nécessaires à cette construction. Cependant, dans le monde animal, nous avons vu que l'exploration d'un environnement inconnu se fait par détection de repères perceptifs, repères utilisés en parallèle pour la construction d'une carte cognitive en temps réel. Pour un robot, résoudre ce problème permettrait de donner à un robot mobile une véritable autonomie dans des environnements inconnus. La **Fig. 2.4** illustre le problème SLAM : un robot parcourt un environnement inconnu, détecte des repères visuels et estime leur position à partir desquelles il va estimer sa propre position. Au temps $t + 1$, et sur la base des informations précédemment perçues, il va (i) mettre à jour les estimations des repères visuels détectés, (ii) sa propre position et surtout (iii) corriger les éventuels imprécisions des estimations de position faites au temps t. L'idée est ainsi de construire, incrémentalement, une représentation interne d'un environnement inconnu à l'aide seulement de ses capteurs (externes ou internes). Ensuite, cette représentation, formalisée par une carte, lui permettra de planifier des mouvements, de naviguer et de se localiser. Cette approche a été un tournant dans les paradigmes d'exploration robotique en cela qu'elle tente de ne se baser que sur les informations perçues par le robot, sans intervention humaine extérieure durant l'exploration. En revanche, préalablement à l'exploration, un grand nombre d'informations sont données au robot, notamment des modèles cinématiques du robot ainsi que des modèles de ses différents capteurs permettant de percevoir l'environnement.

Les premières formalisations mathématiques de l'approche SLAM ont été apportées par Randall C. Smith & Peter C. Cheeseman [40, 41], notamment par l'introduction de la notion de cartes stochastiques qui ont changé la perception de l'incertitude de la mesure des capteurs robotiques en tant que *problème* en un élément *constitutif* de la compréhension du monde perçu par le robot. C'est ainsi que les robots ont commencé à intégrer la notion de probabilité, d'incertitude et d'estimations des diverses grandeurs caractérisant l'environnement. Depuis les années 1990, de très nombreux algorithmes SLAM ont été développés et intégrés à des plateformes robotiques mobiles. Parmi ceux-ci, nous citerons les deux méthodes principales et majoritairement utilisées dans la communauté robotique [42] : l'approche *Extended Kalman Filter* (EKF) et l'approche *FastSLAM*. L'approche EKF-SLAM [43] consiste principalement en l'ajout de bruit blanc gaussien aux modèles dont le robot est doté, le filtre EKF permettant ainsi de résoudre le problème de SLAM. L'algorithme de type FastSLAM (utilisé au sein du projet Two!Ears) a été introduit par Michael Montemerlo *et al.* en 2002 [44] et consiste en considérer le modèle cinématique du robot comme un ensemble d'exemples appartenant à une distribution non gaus-

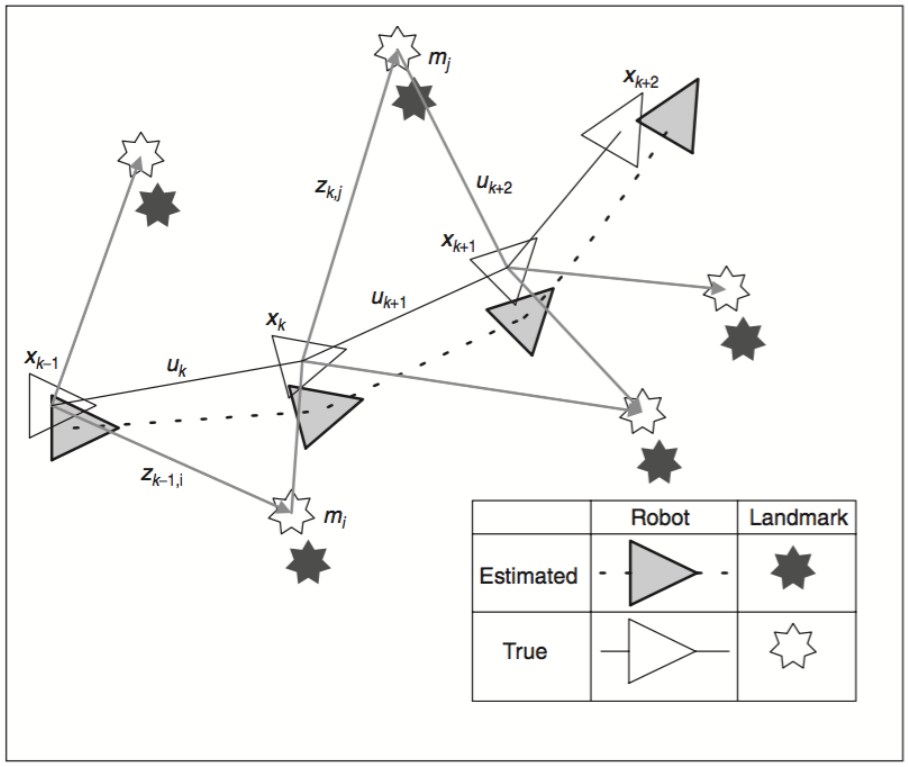


FIGURE 2.4 – LOCALISATION ET CARTOGRAPHIE SIMULTANÉES — Illustration de la problématique de *Simultaneous Localization And Mapping* (SLAM) : une estimation simultanée du robot et de la position des repères est nécessaire afin de permettre à un robot d'explorer et naviguer dans un environnement inconnu de façon autonome. Les positions vraies ne sont jamais connues ou mesurées directement (figure d'après [39]).

sienne, justifiant l'utilisation d'un filtre Rao-Blackwellized [2] [45, 46] pour résoudre le problème SLAM.

Bien que puissants et, d'une certaine façon, bio-inspirés (l'hippocampe est considéré comme effectuant une exploration apparentée au SLAM), les algorithmes SLAM se basent sur une utilisation intensive de modèles donnés au robot avant l'exploration. Nous allons voir, aux sections suivantes, des algorithmes d'exploration presque biomimétiques tentant de reproduire les comportements exploratoires observés chez les animaux par implémentation des cellules responsables de leurs performances, notamment celles décrites à la **Sec. 2.1.1**.

2.1.2.2 Modèles d'exploration bioinspirés

Les modèles présentés dans cette section proposent une implémentation de stratégies d'exploration directement inspirée des processus cérébraux détaillés à la **Sec. 2.1.1**. Une des ambitions des modèles computationnels d'exploration d'environnements inconnus est la création d'une carte cognitive similaire à celle décrite par JOHN O'KEEFE & LYNN NADEL [4] supposément présente dans l'hippocampe des rongeurs. Cette carte congitive, dans le domaine robotique, est censée être une

2. approche bayésienne de l'analyse des données perçues par le robot.

FIGURE 2.5 – INTÉRÊT D'UNE CARTE COGNITIVE DANS LA NAVIGATION — Tâche de navigation dans un environnement volontairement simple : l'agent robotique démarre à la gauche du mur et doit atteindre une position à la droite du mur. (*gauche*) utilisation d'informations visuelles seules, (*centre*) représentation de l'environnement basée sur le principe des *Place cells* créée grâce à une étape d'exploration, (*droite*) combinaison de navigation à l'aide d'informations visuelles et de la carte cognitive précédemment créée (figure d'après [47]).

représentation stable de l'environnement construite itérativement au cours de l'exploration par intégration des différentes sources d'informations auxquelles le système a accès (odométrie, position estimée par les capteurs visuels ou audio etc.). Grâce à elle, il est possible d'améliorer grandement les tâches de navigation et localisation de l'agent robotique au sein de son environnement (cf. **Fig. 2.5**).

Modèle de cellules de transition En 2006 et 2007, NICOLAS CUPERLIER *et al.* [12, 48] ont développé un modèle bio-inspiré basé sur le fonctionnement des *Place cells*, cellules de localisation présentes dans l'hippocampe. Leur ambition est de « développer un système capable de décider de façon autonome un comportement pertinent afin d'accomplir la tâche qui lui incombe ». Les auteurs ajoutent les contraintes suivantes : (i) justification biologique, (ii) modèle minimialiste, et (iii) utilisation de l'information visuelle seulement (pas de laser, ultrasons ou GPS).

De nombreux modèles d'exploration et de cartographie autonome d'un environnement se basent sur les PC du fait de leur caractéristique principale qui est de coder clairement une position spatiale. Ainsi, lorsque l'agent reconnaît un repère visuel, la PC correspondante est activée permettant à l'agent de déduire sa position dans la carte interne qu'il a construite ou qu'il est encore en train de construire. Mais CUPERLIER *et al.* mettent en avant certaines limites dans l'utilisation de ce seul type de cellules, notamment dans des environnements constitués de plusieurs pièces ou lorsqu'une séquence d'actions motrices est nécessaire pour atteindre un point de l'environnement, cas que les PC ne peuvent résoudre à elles seules. Les auteurs proposent donc une modification de l'utilisation des PC inspirée du rôle et du comportement des neurones présents dans le cortex entorhinal (EC) et dans le gyrus denté (DG). Ces structures cérébrales codent la transition spatiotemporelle entre deux PC, au temps t dans l'EC et au temps $t-1$ dans le DG, c'est-à-dire le mouvement nécessaire pour passer d'un point de l'espace à un autre. Ainsi, au lieu de considérer les cellules de position de façon isolée, cette approche tente d'apporter une dimension supplémentaire en intégrant des connexions entre différents points de l'espace. Ces connexions, implémentées sous forme de *cellules de transition* (« *Transition Cells* », TC) dans le modèle de CUPERLIER *et al.* permettent, comme illustré à la **Fig. 2.6**, de coder la séquence d'actions motrices permettant d'aller d'un point

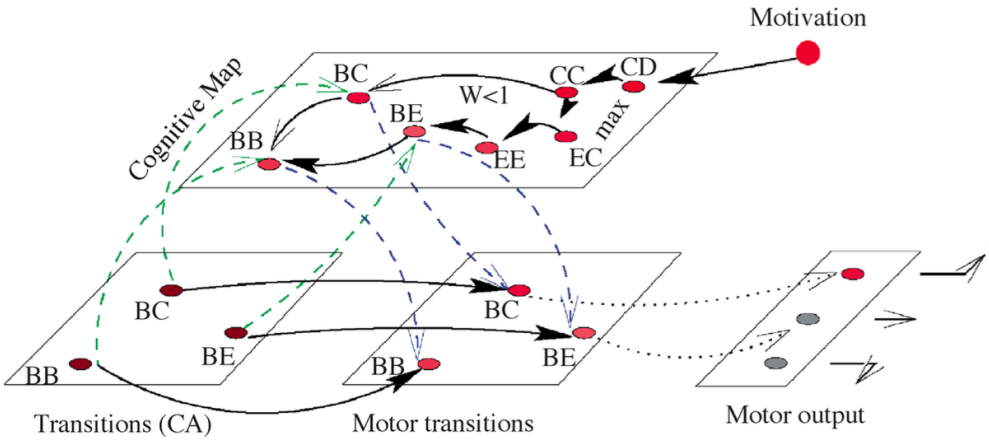


FIGURE 2.6 – MODÈLE DE CUPERLIER *et al.* — Ce modèle implémente les interactions entre l'hippocampe et le cortex préfrontal pour élaborer des stratégies de navigation et de planification en se basant intensément sur l'existence des *Place cells* et des *Transition cells* dans diverses structures cérébrales (figure d'après [12]).

à un autre de l'environnement (les lettres sont des points de l'environnement et les groupes de deux lettres sont les transitions pour aller d'un point à un autre). Sur cette base, une carte cognitive est apprise au cours de l'exploration permettant de naviguer d'un point à l'autre non pas en n'intégrant uniquement la zone d'arrivée à atteindre codée par une PC dédiée, mais en incluant la transition sensorimotrice nécessaire pour passer d'un point à un autre.

Mais la navigation grâce à cette carte est également dépendante d'une entrée cognitive : la motivation à explorer une certaine zone de l'espace. Cette motivation est apportée par la présence dans l'environnement de zones restreintes possédant un intérêt éventuel pour le robot : un point de nourriture ou une zone de repos, par exemple. De façon similaire aux motivations à l'exploration présentées à la **Sec. 2.1.3**, l'exploration du robot soumis au modèle de CUPERLIER *et al.* pourra être spécifiquement motivé par la recherche d'une zone de nourriture par exemple, ajoutant une contrainte dans le choix de la cellule de transition gagnante et donc dans la prochaine séquence d'actions motrices que le robot va devoir suivre.

Cette approche permet la construction en ligne d'une carte cognitive de l'environnement divisée en régions distinctes et permettant aux robots sur lesquels le modèle a été testé d'effectuer des tâches de navigation et de cartographie de façon pertinente. L'intérêt de ce modèle réside tant dans son inspiration biologique que dans sa relative simplicité. En effet, se basant uniquement sur les cellules codant une position de l'espace (les PC) et celles liant deux PC entre elles (les TC), le modèle parvient à faire émerger un comportement exploratoire efficace. De plus, l'intégration d'un « moteur motivationnel » (notion détaillée plus tard) constitue un exemple de l'intérêt d'ajouter des concepts empruntés à la cognition, même bas-niveau, dans les stratégies d'exploration d'un environnement et de navigation dirigée, au sein de celui-ci. Dans la même veine que ce modèle, l'approche AniMat tente de modéliser le comportement exploratoire observé chez les animaux, les rongeurs notamment, mais en allant parfois plus loin, jusqu'au biomimétisme.

RatSLAM (Animat) En 2004, Michael Milford *et al.* [49] ont proposé une nouvelle approche du paradigme SLAM (cf. **Sec. 2.1.2.1**) nommée *RatSLAM* et directement inspirée par l'hippocampe des rongeurs. Sur la base d'informations allothétiques (capteurs visuels) et idiothétiques (odométrie), l'approche RatSLAM est une tentative de modélisation de l'hippocampe *via* l'utilisation de réseaux de neurones mimant l'activité des cellules de cette structure cérébrale. L'idée est de coder la *pose* du robot, définie comme la direction angulaire de la tête θ et la position spatiale du robot (x, y) dans l'environnement. Le choix de ces deux types d'informations est une nouvelle fois motivé par des observations biologiques : la direction angulaire est une donnée codée par les cellules dédiées *Head Direction cells* mises en avant chez le rongeur notamment, la position spatiale étant notamment codée par les *Place cells*. La conjonction des deux est appelé une *cellule de pose* (« *Pose Cell* » PoC). De précédents systèmes computationnels ont séparé ces deux informations en deux réseaux de neurones distincts ayant pour limite l'impossibilité de représenter et de maintenir de multiples hypothèses sur la pose au cours du temps. Le modèle RatSLAM quant à lui tente de combiner angle de la tête et position spatiale au sein d'un seul réseau afin de rendre possible l'émission de multiples hypothèses sur la pose du robot.

Les résultats obtenus par Milford *et al.* sur un vrai robot montrent l'intérêt d'ajouter l'information sur la direction de la tête dans l'exploration d'un environnement. L'algorithme RatSLAM permet de donner à un robot une bonne capacité à créer une carte cognitive d'un environnement inconnu, de façon autonome et suffisamment robuste pour pouvoir s'y déplacer correctement. Cet algorithme surpasse même dans certaines conditions les approches conventionnelles de type SLAM. A noter qu'en 2006, Milford *et al.* [50] ont apporté une amélioration au paradigme RatSLAM en donnant à l'algorithme la capacité de « désapprendre » certaines routes qu'il avait précédemment apprises, *via* l'introduction d'une « carte d'expérience ». Cette carte additionnelle permet notamment de moduler le comportement des cellules de pose. L'adaptabilité conférée par la carte d'expérience se révèle extrêmement utile lorsque des obstacles apparaissent dans des zones de l'environnement déjà explorées. Grâce à elle, l'algorithme RatSLAM devient moins sensible aux variations de l'environnement déjà exploré.

Mais si les modèles de Cuperlier *et al.* et de Milford *et al.* se basent intensivement sur l'hippocampe, le modèle présenté ci-après, *PsiKharpax*, a été directement inspiré du rôle des ganglions de la base, structure cérébrale impliquée dans le choix de la prochaine action à réaliser, mécanisme majeur dans le cadre de l'élaboration d'une stratégie d'exploration.

PsiKharpax (Animat) Le projet PsiKharpax a été initié par Jean-Arcadi Meyer *et al.* en 2005 [51] avec pour ambition la conception de capteurs biomimétiques et d'architectures de contrôles neuronaux permettant de doter un robot de capacités d'autonomie et d'adaptation. Les capteurs allothétiques de PsiKharpax sont une paire d'yeux, une paire d'oreilles (cochlées) et soixante-quatre vibrisses. Les capteurs idiothétiques sont un système vestibulaire (permettant de calculer les accélérations linéaires et angulaires de la tête), un système odométrique (permettant la surveillance de la longueur et de la direction des déplacements), et un système de vérification de l'état énergétique du robot. D'autre part, la tête dont PsiKharpax est

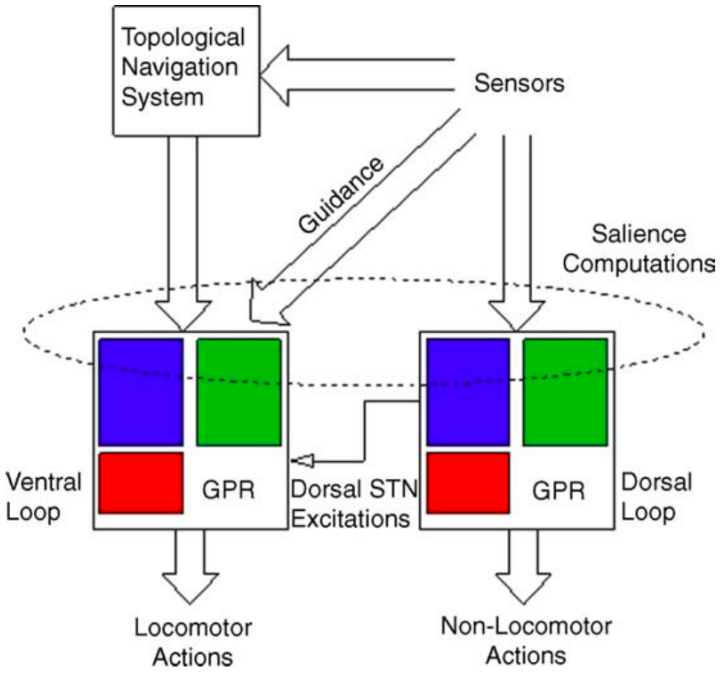


FIGURE 2.7 – PSIKHARPAX & MODÈLE GPR — Schéma simplifié du modèle mettant en avant l'utilisation d'unités de type GPR (décrites à la section suivante) permettant de résoudre le problème de sélection de la prochaine action à exécuter dans les tâches d'exploration (figure d'après [51]).

dotée est capable de tourner, ajoutant ainsi une forme d'exploration visuelle supplémentaire. De plus, des réflexes bas-niveau sont implémentés comme, par exemple, le suivi d'un objet lorsque la tête est en mouvement ou l'évitement d'obstacles détectés par les moustaches de PsiKharpax ou par son système visuel ou auditif.

Le modèle de navigation de PsiKharpax est basé sur une stratégie de traque d'objet à multiples hypothèses inspirée par les PC et les HDC, chacune de ces hypothèses étant mise à jour en parallèle et aboutissant à l'élaboration d'une carte topologique. Cette carte est composée de nœuds qui codent les données allothétiques perçues par le robot. Chaque nœud est lié aux autres nœuds de la carte, le lien entre deux nœuds codant la distance entre eux ainsi que leur position relative, en fonction des données idiothétiques issues des capteurs du robot. L'activité de chaque nœud représente la probabilité que PsiKharpax soit à cette position.

L'exploration d'un environnement inconnu par PsiKharpax a, de plus, été agrémentée de tâches possibles à effectuer dans un contexte de *survie* : exploration sans but (« *wandering* »), évitement d'obstacles, nourriture (i.e. énergie du robot) et repos (lorsque le niveau d'énergie est trop bas). Le robot doit seul être capable de déterminer quelle est la tâche à effectuer en fonction des données idiothétiques auxquelles il a accès, notamment son niveau d'énergie. L'intérêt de l'approche PsiKharpax réside dans l'ajout d'une modélisation de l'action des ganglions de la base dans le processus de sélection de l'action, telle que formalisée par GURNEY *et al.* dans le modèle GPR (décrit à la section suivante). Les actions de PsiKharpax sont ainsi calculées grâce à ces composants neuromimétiques GPR, comme illlustré à la **Fig. 2.7**.

En plus de la boucle de rétrocontrôle du modèle GPR, deux boucles additionnelles

ont été implémentées, inspirées par l'organisation des flux perceptifs et analytiques cérébraux : une boucle *ventrale* qui permet de sélectionner une action locomotrice, et une boucle *dorsale* qui sélectionne les actions non-locomotrices, les deux étant également modélisées par un système de type GPR. Les connexions entre ces deux boucles empêchent le robot d'effectuer une action locomotrice et non-locomotrice en même temps. En effet, la boucle dorsale envoie des entrées excitatoires à la boucle ventrale entrainant une augmentation du niveau d'inhibition de toutes les actions motrices.

En fonction de la motivation à court-terme du robot, différents profils de navigation peuvent être sollicités :

- *Planning* : profil correspondant à la planification d'une navigation motivée par l'attraction vers deux sources connues et situées à des positions différentes de l'environnement ;
- *Homing* : profil correspondant à la motivation à réexplorer des régions de l'environnement déjà connues ;
- *Exploration* : profil correspondant à la motivation à l'exploration de régions inconnues

Pour résumer les capacités de PsiKharpax, nous citerons Meyer *et al.* [51] :

> « *PsiKharpax, by being able to integrate the past (through its recorded map), the present (through its sensors) and the future (through its planning capacities), will represent an embodied example of a motivationally autonomous animat whose control complexity may well challenge the possibilities of external control and, hence, its capacities to withstand any imposed autonomy*[3]*.* »

Modèle GPR Le modèle **G**urney **P**rescott **R**edgrave (GPR [52, 53]) a pour but de modéliser l'activité des noyaux des ganglions de la base selon la boucle *ganglions de la base — thalamus — cortex*. L'idée est que chaque action motrice discrète est codée dans des canaux d'information distincts, situés dans les noyaux des ganglions de la base. Ces canaux portent l'information sur les actions motrices possibles et sont, par défaut, inhibés. Les ganglions de la base vont alors collecter ces afférences, les intégrer et sélectionner le canal le moins inhibé afin de promouvoir l'action motrice qu'il représente. Les entrées de ces canaux sont appelés des *saillances* incluant des perceptions autant internes qu'externes (données allothétiques et idiothétiques), afin d'évaluer quelle action est la plus pertinente en fonction des besoins du robot (cf. **Fig. 2.8**). De plus, une boucle de rétrocontrôle positive entre le thalamus et ces saillances induit une persistence dans le choix des actions. Au final, la décision est prise selon l'action qui aura été la moins inhibée.

Le modèle GPR est intéressant en cela qu'il permet de doter des algorithmes d'exploration (entre autres) d'une structure bio-inspirée modélisant le phénomène de

3. « PsiKharpax, par sa capacité à intégrer le passé (grâce à sa carte interne du monde), le présent (grâce à ses capteurs) et le futur (grâce à ses capacités de planification), représentera un exemple incarné d'Animat autonome et doté de motivation et dont la complexité de contrôle pourrait concurrencer les algorithmes de contrôle externes et, ce faisant, ses capacités à résister à n'importe quelle autonomie imposée. »

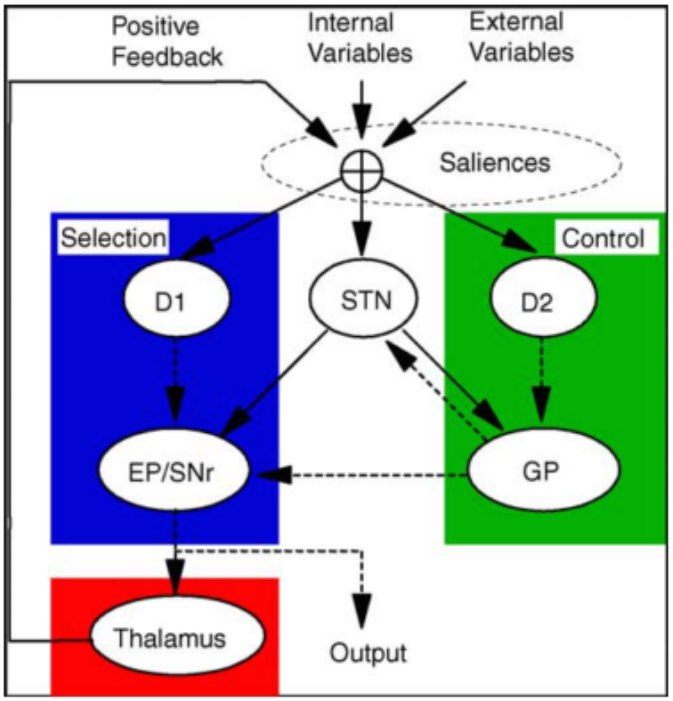


FIGURE 2.8 – MODÈLE GPR — Illustration d'un des canaux des ganglions de la base selon le modèle GPR. (*flèches pleines*) connexions excitatoires, (*flèches pointillées*) connexions inhibitrices (figure d'après [51]).

sélection de l'action tel qu'il a été observé au niveau des ganglions de la base. Dans le cadre du modèle HTM, nous allons utiliser le modèle GPR et son inspiration biologique pour déterminer vers quelle source audiovisuelle le robot doit tourner sa tête, notamment lorsqu'il y a plusieurs candidats possibles.

2.1.2.3 Discussion

Nous avons brièvement présenté dans cette section quelques-uns des très nombreux modèles d'exploration inspirés par des structures cérébrales identifiées chez le rongeur aussi bien que chez l'homme, notamment l'hippocampe, le cortex entorhinal et les ganglions de la base. Que ce soit par l'utilisation de *Place cells*, de *Transition cells*, des canaux d'informations de type GPR ou par l'élaboration de cartes cognitives mettant en relation l'agent robotique et son environnement — sur la base de repères généralement visuels — les modèles présentés ici confèrent aux robots dans lesquels ils ont été intégrés la possibilité d'explorer des environnements inconnus de façon rapide et performante tout en aboutissant à une représentation interne de ces environnements similaire à celles observées chez l'animal. Cela nous permet d'entrevoir à quel point l'efficacité de l'architecture du cerveau peut permettre l'élaboration de systèmes computationnels d'exploration performants tout en conservant une certaine simplicité dans leur conception. Le modèle HTM tire également profit des travaux de recherche menés sur ces structures cérébrales en cela qu'il aboutira à une forme de carte cognitive constituée des objets audiovisuels perçus au cours de l'exploration. De plus, le fonctionnement de la boucle ganglions de la base — thalamus — cortex et particulièrement son implémentation en modèle

GPR nous ont inspiré pour le mécanisme de sélection de l'origine des mouvements de tête effectué par le modèle HTM (décrits plus tard). En effet, les deux modules constitutifs du modèle HTM sont capables de générer des mouvements de tête et leur décision peut être contradictoire. Ainsi, afin de décider quel module prend le pas sur l'autre, l'algorithme que nous avons développé s'est inspiré du principe des canaux d'informations inhibés par défaut et dont la valeur de l'activité permet la prise de décision sur le module gagnant.

Les modèles présentés ici ont pour *but* de cartographier l'environnement. Même si certains d'entre eux incluent déjà une forme de motivation (le modèle de Cuperlier *et al.*), la seule forme de « motivation » prise en compte est celle de l'exploration la plus exhaustive de l'environnement. Or il existe de nombreuses autres sources de motivations à explorer un environnement inconnu, moivations permettant d'accomplir ces tâches exploratoires selon d'autres mécanismes et d'autres buts globaux à accomplir. La section suivante introduit donc un autre pan de la modélisation de comportements exploratoires centrée autour de ces buts plus haut-niveau conditionnant l'exploration d'environnements inconnus ainsi que l'interaction d'un agent robotique au sein de ceux-ci.

2.1.3 Motivations pour l'exploration

Depuis des décennies, la *Motivation* a été considérée comme un mécanisme fondamental pour expliquer les comportements exploratoires spontanés chez l'humain, et chez l'enfant en particulier [54]. L'exploration est autant influencée par des caractéristiques intrinsèques des stimuli perçus telles l'intensité, la couleur ou la hauteur du son, que par des éléments plus cognitifs tels les notions de récompense ou de punition. Cependant, des principes plus haut-niveau et à portée plus générale sont également déterminants dans l'exploration d'un environnement : la *Nouveauté*, le *Changement*, la *Surprise* ou encore la *Curiosité*. Bien que ces principes pourraient être considérés comme des émotions, ils sont plutôt définis comme des « moteurs motivationnels ». En effet, une distinction doit être faite entre (i) une stratégie ayant pour but de rechercher les événements perceptifs qui causeront le *sentiment de surprise/changement/curiosité* et (ii) la réaction biologique provoquée par des stimuli nouveaux, incongrus ou inconnus.

Durant les vingt dernières années, plusieurs types de motivations ont été conceptualisées, au point qu'elles ont été implémentées dans plusieurs plateformes robotiques, simulées ou réelles. Les motivations à explorer peuvent différer des générateurs de buts classiques comme, par exemple, l'exploration complète d'un environnement inconnu ou la recherche du prochain meilleur point d'observation. La notion de *Motivation* se place en amont de ces paradigmes traditionnels d'exploration/action en cela qu'une notion de *récompense* y est ajoutée. C'est cette récompense qui est supposée fournir au robot une satisfaction *intrinsèque* suffisamment forte pour lui apporter la motivation à explorer son environnement. Tout l'enjeu de ces approches est donc de pouvoir formaliser la manière dont le système artificiel va trouver de l'« intérêt » à effectuer une tâche.

En 2000, Richard M. Ryan & Edward L. Deci [55], inspirés par les travaux précurseurs de Daniel Ellis Berlyne en 1950 [56], définissent la motivation intrin-

sèque comme le fait de « réaliser une action pour sa satisfaction inhérente plutôt que pour une conséquence séparable [4] ». La motivation intrinsèque peut ainsi être définie comme une activité plaisante pour le robot, à l'opposé de la motivation extrinsèque qui est l'intérêt d'accomplir une tâche pour une entité exterieure, l'expérimentateur humain notamment. Récemment, de nombreuses tentatives de modélisation et d'intégration de ces modèles dans des robots exploratoires ont été guidées par cette notion de motivation intrinsèque (voir [57, 58, 59]). L'exploration guidée par une motivation intrinsèque n'est pas homéostatique : le désir d'explorer un environnement n'est pas causé par un simple *besoin* auquel il faut répondre dans le but de conserver le système dans un état énergétique stable et constant. Par exemple, la perturbation biologique causée par l'apparition inattendue d'un nouveau stimulus peut être la réponse à un besoin : celui de diminuer l'incertitude passée ayant causée cette réaction. Au-delà de ce « besoin », une notion de « désir », qui est, par définition, intrinsèque, peut motiver à réduire cette incertitude. Selon la définition de Ryan & Deci de la motivation intrinsèque, Pierre-Yves Oudeyer propose une définition computationnelle nouvelle :

> « *An experienced situation [...] is intrinsically motivating for an autonomous entity if its interest depends primarily on the collation or comparison of information from different stimuli and independently of their semantics, whether they be physical or imaginary stimuli (i.e. measured by physical sensors or by internal 'software' sensors) perceived in the present or in the past* [5]. »

Ici, la notion d'information est comprise selon une approche informationnelle théorique définie par la structure mathématique intrinsèque des stimuli et non par le sens qu'ils portent. Un système prenant en compte la motivation intrinsèque doit donc intégrer un mécanisme capable d'évaluer la propension d'une situation à évoquer la surprise, la complexité, le *challenge* ou la nouveauté, pour le robot, tout en y associant une récompense. Maximiser ces mesures peut aboutir à l'élaboration d'une exploration active et autonome. La motivation intrinsèque regroupe plusieurs types de motivations, telles que :

- la *motivation par l'incertitude*, définie comme l'attraction pour de nouveaux stimuli. Ainsi, pour chaque événement observé, une récompense va être générée, de façon inversement proportionnelle à sa probabilité d'observation [60] ;
- la *motivation par le gain d'information*, pouvant être définie comme le « plaisir d'apprendre » et qui pousse le robot à minimiser le niveau d'incertitude dans la connaissance qu'il a de l'environnement [61]. Ce type de motivation peut être différente de la motivation par l'incertitude : une zone incertaine de l'environnement peut être définie comme contenant probablement peu de nouvelles informations ;

4. « *The doing of an activity for its inherent satisfaction rather than for some separable consequence* »

5. « Une situation perçue [...] est intrinsèquement motivante pour une entité autonome si son intérêt dépend principalement de la comparaison de l'information portée par différents stimuli, indépendamment de leur sémantique et qu'ils soient physiques ou imaginaires (c'est-à-dire des stimuli perçus par les capteurs physiques ou par des capteurs *logiciels* internes) perçus dans le présent ou le passé. »

- la *motivation de l'autonomisation*, définie comme la recherche de l'acquisition du maximum d'information par les capteurs du robot. Ainsi, le robot va tenter de trouver la séquence d'actions produisant le flux le plus important d'informations [62].

De façon plus concrète, les principales motivations intrinsèques permettant l'émergence de stratégies d'exploration sont la *Curiosité*, la *Surprise* et la *Faim* notamment. La première peut être définie comme le désir d'aquérir de l'information sur tout nouvel objet ou sur un objet dont certaines caractéristiques restent incertaines mais qui semblent d'intérêt [56]. Il s'agit d'une motivation à réduire l'incertitude sur l'environnement par exploration de zones inconnues. De plus, le manque d'information sur des objets présents dans ces environnements sont également susceptibles de motiver un robot dirigé par la Curiosité à aller les explorer. La *Faim* est le simple besoin de trouver une source d'énergie. Bien que la Faim puisse être vue comme un véritable besoin, notamment dans les scénarios robotiques de « survie » dans lesquels des paramètres tels que le niveau de batterie est géré de façon autonome par le robot, il est également possible de la considérer comme une motivation si la recherche de zones dans lesquelles des sources d'énergie se trouvent est associée à une récompense plutôt qu'au danger auquel le robot fait face (celui de s'éteindre). Les quelques modèles présentés dans cette section permettent d'entrevoir les travaux menés en robotique exploratoire qui ont intégré ce concept de motivation.

2.1.3.1 Le modèle de Schmidhuber

En 1991, Jürgen Schmidhuber [63] a implémenté un agent artificiel doté de Curiosité et d'Ennui, un des premiers dans ce domaine de recherche. Ce modèle tente de mettre l'agent dans des situations pour lesquelles il aura le plus de chances d'apprendre quelque chose de nouveau sur l'environnement. L'idée est d'apprendre à estimer les effets d'une poursuite de l'apprentissage *versus* son arrêt sur la capacité du robot à se représenter l'environnement. Pour cela, deux modules constituent le modèle de Schmidhuber : un module de « confiance adapatative » et un module de « curiosité adaptative ». Le premier permet de mesurer à quel point l'agent robotique est confiant en sa perception de l'environnement tandis que le second permet de motiver l'exploration de zones mal connues. Ce modèle doté de curiosité a permis d'améliorer grandement l'exploration d'un environnement inconnu en comparaison d'une exploration aléatoire. Il consitue une des premières tentatives de formalisation de moteurs motivationnels intégrés à un robot.

2.1.3.2 Le modèle de Macedo

Au début des années 2000, Macedo *et al.* [64, 5] ont étudié le rôle et l'importance respective des motivations suivantes : Curiosité, Faim et Surprise. En particulier, en 2005 [65], ils ont implémenté un module motivationnel permettant de passer d'un type de motivation à l'autre et d'observer l'impact sur un même scénario d'exploration (cf. **Fig. 2.9**). Le système motivationnel implémenté génère des buts et des intentions données au robot évoluant dans des environnements simulés. Couplé à ce système, un module de délibération et de prise de décision va combiner ces

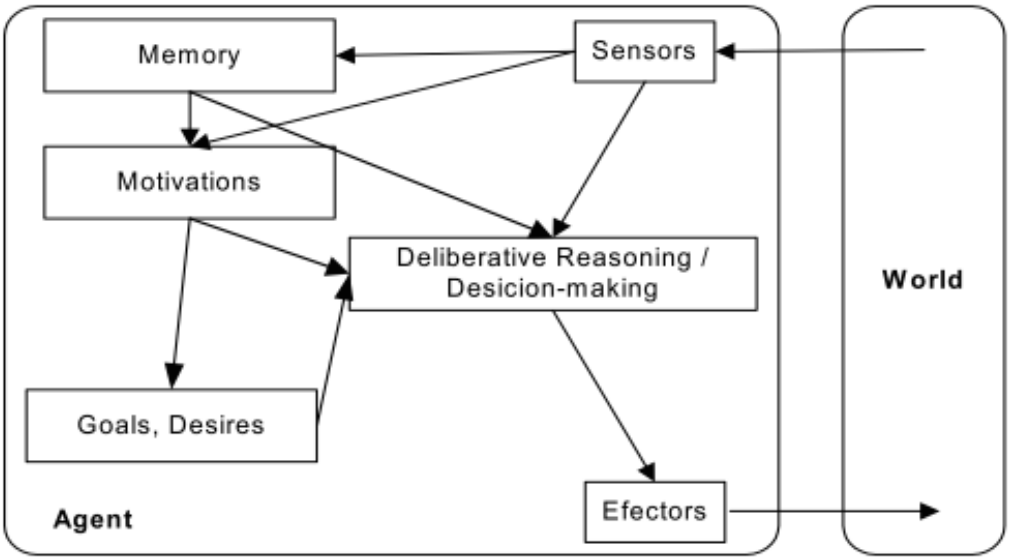


FIGURE 2.9 – MODÈLE DE MACEDO *et al.* — Architecture du système permettant de tester l'impact sur l'exploration d'un environnement inconnu des motivations intrinsèques suivantes : *Curiosité*, *Surprise* et/ou *Faim* (figure d'après [65]).

sources de motivation et éventuellement les pondérer, en fonction du scénario considéré. D'autre part, leur approche se base sur une conception assez haut niveau de l'environnement : en addition aux mesures et cartes topologiques du terrain exploré, le robot simulé interprète son environnement notamment par le biais des objets qui le composent. La carte cognitive créée lors de l'exploration, forme de mémoire épisodique et sémantique, sert évidemment à la navigation mais également à la recherche de zones offrant une récompense particulière, en fonction de la motivation qui aura été conférée au robot.

Les résultats montrent que la *Faim*, motivation à trouver une source d'énergie, semble être le moteur motivationnel le plus puissant afin d'effectuer une exploration exhaustive et rapide de l'environnement mais que la *Surprise* confère une exploration plus rapide, bien que moins performante. Malgré ce résultat, assez isolé, la plupart des travaux de recherche de ces quinze dernières années se sont penchés sur la *Curiosité* en tant que motivation entrainant une exploration puissante par réduction de l'incertitude sur l'environnement.

2.1.3.3 Les modèles d'Oudeyer et Baranes

ANDREW G. BARTO *et al.*, en 2004 [66], sont parmi les premiers à avoir proposé une formalisation *mathématique* de la motivation intrinsèque chez un robot. Bien que l'application de leur formalisation de la motivation intrinsèque ne soit pas exclusivement destinée à l'exploration d'un environnement inconnu, il a servi de base algorithmique à de nombreux autres modèles, destinés, eux, à des tâches d'exploration comme notamment l'algorithme *Intelligent Adaptive Curiosity* [67] (IAC). Un des fondements de l'algorithme IAC est l'apprentissage par renforcement, paradigme d'apprentissage machine basé sur la notion de *récompense*, la plupart du temps liée à la réalisation d'une action. La motivation employée ici est le maintien

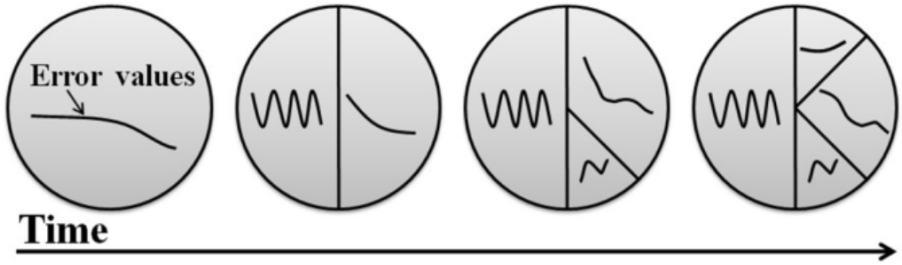


FIGURE 2.10 – ALGORITHME R-IAC — Illustration du principe de découpage de l'environnement sensorimoteur sur la base du modèle IAC [66]. Le principe est de découper la représentation de l'environnement en différentes régions maximisant leur différences respectives afin de motiver l'exploration la plus efficace résultant en la diminution rapide des erreurs de prédiction de chaque région (figure d'après [58]).

d'un état cognitif de l'agent : son degré d'apprentissage (« learning progress »), qui reflète la connaissance que le robot a de son environnement. Ce degré d'apprentissage doit rester maximal afin de réduire, autant que possible, l'incertitude sur certaines zones de l'environnement. L'idée est proche de celle de *Curiosité* en cela que l'algorithme pousse le robot a expérimenter de nouvelles situations afin d'engranger de nouvelles informations participant à la construction de sa représentation interne de l'environnement. L'« intelligence » de cet algorithme que mentionne OUDEYER *et al.* provient de la capacité de cette approche à éviter de se trouver piégé dans deux cas extrêmes : une situation trop *prédictible* ou, à l'inverse, une situation trop *imprédictible*, ces deux situations étant également mentionnées comme « *the edge of order and chaos* » dans les théories de dynamique cognitive. Le fonctionnement d'IAC est le suivant : une représentation sensorimotrice des expériences passées du robot est découpée en *régions* codant des ensembles d'exemples perceptifs que le robot a vécu [6]. A chaque région est couplé un algorithme d'apprentissage (réseau de neurones, support vecteur-machine, machine bayésienne etc.) dont le but est d'apprendre à prédire un ensemble de données sensorimotrices au temps $t+1$ sur la base des données de la région concernée au temps t. Enfin, l'attraction d'une région de l'espace sensorimoteur, qui servira pour la décision de la prochaine action motrice à effectuer, sera calculée selon la qualité de la prédiction faite par chacun des algorithmes d'apprentissage (un par région). Le réseau le moins performant sera celui qui présentera le plus grand taux d'erreur de prédiction et représentera une récompense d'autant plus forte. C'est cette récompense qui motivera l'exploration de la région de l'espace correspondant à ce réseau le moins performant.

Les résultats obtenus par OUDEYER *et al.* montre l'émergence de comportements exploratoires performants et pertinents dans des conditions simulées aussi bien que dans des environnements réels, bien que simples. Cette approche, basée sur un apprentissage en ligne de l'expérience passée du robot et y couplant une intégration multimodale (sensorimotricité) aboutit à une compréhension riche de l'environnement.

Sur cette base, ADRIEN BARANES & PIERRE-YVES OUDEYER, en 2009 [58] ont

6. A noter que la façon dont le découpage de l'espace sensorimoteur en régions se déroule est déterminé par deux seuils définis préalablement par l'expérimentateur.

proposé l'algorithme R-IAC (pour *Robust-IAC*) permettant d'améliorer l'étape de découpage de l'espace sensorimoteur en différentes régions. Les auteurs incorporent une forme d'apprentissage progressif grâce auquel la division d'une région de l'espace en deux régions filles n'est plus faite de façon neutre mais de telle sorte que la dissimilarité entre ces deux nouvelles régions soit maximale, afin d'encourager le système à apprendre des situations différentes (cf. **Fig. 2.10**). Cette modification substantielle améliore grandement les résultats de l'exploration d'un espace du point de vue sensorimoteur.

L'année suivante, les auteurs ont proposé une amélioration à l'algorithme R-IAC en l'implémentation de l'algorithme *Self-Adaptive Goal Generation Robust-Intelligent Adaptive Curiosity* [59] (SAGG-RIAC). L'algorithme SAGG-RIAC a été développé dans selon le concept de « Competence Based Active Motor Learning »[7] [57] consistant en une architecture à deux niveaux aux dynamiques temporelles différentes (cf. **Fig. 2.11**) :

échelle temporelle courte, prenant en compte les actions bas-niveau à effectuer afin d'accomplir un but déterminé dans l'exploration active dépendant de mesures locales de l'évolution de la qualité de l'apprentissage des modèles inverses et/ou *forward* ;

échelle temporelle longue, prenant en compte les buts générés et sélectionnés par le système lui-même, buts dépendants de flux ascendants et descendants d'informations sur le niveau d'achèvement du but précédemment généré.

La conception de mécanismes d'exploration et d'apprentissage *goal-directed* (échelle temporelle courte) inclut un modèle inverse et/ou *forward* généré durant l'exploration et disponible pour une utilisation future, ainsi qu'un retour de l'avancement et de la qualité de cet apprentissage permettant l'élaboration de nouvelles actions dans la tâche d'exploration active. Le processus de self-génération et self-sélection de buts (échelle temporelle longue) est basé sur l'amélioration de la compétence dans des sous-régions de l'espace dans lesquelles les buts sont choisis [59]. Cette notion de *Compétence* est la motivation intrinsèque du robot.

2.1.3.4 Discussion

Dans cette section nous avons détaillé le concept de *Motivation* à explorer un environnement, notion notamment introduite et décrite par Daniel Ellis Berlyne en 1950. Centrer un modèle d'exploration robotique autour d'une telle notion place les systèmes computationnels créés à un plus haut-niveau : celui de la cognition. L'introduction de récompense et de but à long terme confère en effet à un robot mobile une compréhension plus large et plus complexe de son environnement — et de lui-même — qu'une exploration basée exclusivement sur la création d'une carte cognitive purement topologique.

Le modèle IAC (et ses extensions : R-IAC et SAGG-RIAC) a été une forte source d'inspiration lors de la conception du modèle HTM en cela que le principe de lier une récompense élevée à la région de l'espace la moins bien représentée, du point de vue du robot, a été similairement utilisée lors de l'implémentation d'une partie

7. Apprentissage moteur actif basé sur la compétence.

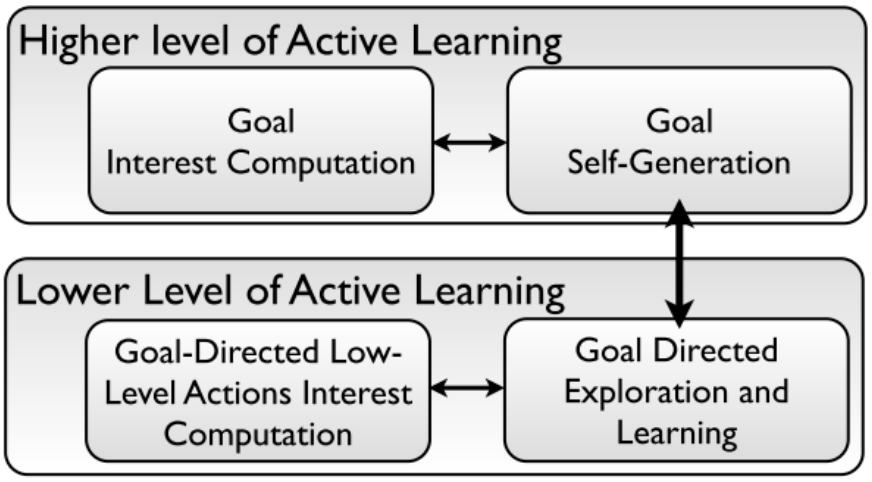


FIGURE 2.11 – ALGORITHME SAGG-RIAC — Deux parties composent cet algorithme, chacune définissant deux niveaux d'apprentissage actif : une partie haut-niveau impliquée dans la self-génération et la self-sélection active de buts à accomplir ; une partie plus bas-niveau impliquée dans la sélection d'actions bas-niveau (telles que des actions motrices) dirigées par les contraintes imposées par la partie haut-niveau (figure d'après [59]).

---

du modèle (décrit au **Chap. 6**). De plus, l'association sensorimotrice est un exemple d'intégration de données multimodales dans un cadre d'apprentissage par renforcement visant à obtenir une représentation de l'environnement riche et robuste aux variations de la qualité et/ou de l'accessibilité des données perçues par le robot.

Le modèle HTM quant à lui, par sa capacité à générer des commandes motrices de type rotation de tête, est un modèle conférant au robot la capacité d'explorer son environnement. Il est constitué de deux parties (deux modules) ayant chacun leur propre motivation à déclencher ces mouvements de tête : l'un est motivé par la *Suprise*, l'autre par la réduction de l'*Incertitude*, forme de *Curiosité*. De plus, le paradigme d'apprentissage utilisé par le module MFI est similaire à celui formalisé par BARTO *et al.* en cela qu'il correspond à un découpage d'un espace multimodal (sensorimoteur dans leur cas, audiovisuel dans le nôtre) en différentes régions, chacune étant capable de coder plus ou moins correctement l'information qu'elle est censée représenter.

2.1.4 Conclusion

Cette section dédiée à l'exploration a tenté d'offrir une vue d'ensemble du domaine de l'exploration, commençant par ses structures cérébrales et se poursuivant par leur modélisation et souvent leur implémation dans des plateformes robotiques. L'accent a été mis sur les travaux de recherche admettant une inspiration biologique claire. Plus qu'une inspiration, il s'agit d'un choix philosophique conduisant à l'élaboration de paradigmes utilisés lors de la création de systèmes artificiels visant à reproduire les mécanismes observés chez les animaux. Nous nous situons dans cette même veine, considérant que les connaissances issues de l'observation des mécanismes biologiques est une source intarissable d'idées, de concepts et d'architectures

dont nous pouvons juger de leur incroyable performance.

Cette section a tout d'abord exposé les principales structures cérébrales impliquées dans la cartographie d'un environnement inconnu et la navigation ultérieure grâce à cette carte. L'hippocampe, et sa connexion avec le cortex entorhinal, semble être la structure qui rassemble les différents processus conduisant à l'élaboration d'une telle carte cognitive, bien que sa seule implication ne semble pas être suffisante pour expliquer l'ensemble des phénomènes — et leur complexité — permettant à un animal de se repérer dans un environnement, connu ou non. La présence de cellules très spécialisées dans le codage des informations spatiales, comme les *Place cells*, les *Grid cells* ou les *Head direction cells*, ainsi que l'intégration de différentes sources d'information (vision, audition, odorat etc.) rend ces structures capables de créer des représentations internes diverses et très adaptatives des environnements explorés. De plus, les ganglions de la base, par leur rôle dans les mécanismes de sélection de l'action, permettent, conjointement à l'utilisation des cartes cognitives créées, de raffiner le processus d'exploration motrice.

Sur cette base, la communauté robotique bio-inspirée a développé de très nombreux algorithmes d'exploration. Bien que, du point de vue ingéniérie, les techniques d'exploration répondant au problème général de cartographie et localisation simultanée *Simultaneous Localization And Mapping* (SLAM) ne soient pas forcément considérées comme bio-inspirées, il est important de noter que le simple fait de tenter d'élaborer des algorithmes capables de conférer à un robot la capacité à créer une carte d'un environnement inconnu en quasi temps réel afin de pouvoir s'y déplacer de façon autonome quasi instantanément, est déjà une inspiration des comportements exploratoires animaux très forte. L'algorithme FastSLAM mentionné plus haut et utilisé au sein du projet Two!Ears se sert notamment d'un filtre basé sur une approche bayésienne, approche basée sur un paradigme également largement observé dans de nombreux mécanismes cérébraux, notamment dans la perception.

Les modèles biomimétiques quant à eux, notamment ceux entrant dans la catégorie des robots de type Animat, comme RatSlam ou PsiKharpax, ont tous comme point commun leur inspiration de structures cérébrales comme l'hippocampe, le cortex entorhinal ou les ganglions de la base, parfois même une combinaison complexe des trois. Et au-delà de la simple modélisation du comportement de structures cérébrales, ces approches sont implémentées dans des robots aux capteurs également biomimétiques. Notamment, PsiKharpax est doté d'une vision binoculaire, d'une audition binaurale et est capable de bouger sa tête, caractéristiques que le robot sur lequel le modèle HTM a été intégré partage également.

Mais en parallèle de ces tentatives de modélisation de la capacité à transformer des informations spatiales perçues au cours de l'exploration d'un environnement en une carte stable et réutilisable, un pan entier de la recherche robotique mobile et comportementale a tenté de définir les motivations qui, en amont, poussent un animal à explorer son environnement. Ces motivations, sur la base des définitions de Berlyne, peuvent changer la façon dont un agent robotique considère son environnement ainsi que les besoins auxquels il tente de répondre. Cette approche, plus cognitive, tente d'introduire des notions telles que la *Surprise*, la *Faim* ou la motivation par la réduction de l'*Incertitude*. Ces travaux ne sont pas antinomiques avec les algorithmes d'exploration cités plus haut, bien au contraire. Car d'un côté se situe la

façon dont un robot collecte des informations spatiales et les ordonne afin d'en tirer une représentation cohérente de l'environnement et de l'autre se situe le but global de cette exploration, but modifiant le choix de la prochaine zone de l'environnement à explorer.

Le modèle HTM s'inspire et s'inscrit à la suite de ces travaux en cela qu'il constitue une *motivation pour explorer un environnement inconnu.* Nous définissons l'environnement, dans le cadre du modèle HTM, comme l'*ensemble des entités audiovisuelles qui le composent.* Reconsidérant la **Fig. 2.3** schématisant les différentes tâches impliquées dans le domaine de l'exploration robotique, le modèle HTM est impliqué dans les tâches de cartographie et de contrôle moteur : cartographie par détection de sources audiovisuelles et localisation de celles-ci dans le référentiel interne au robot, contrôle moteur en cela que le modèle gère les mouvements de tête du robot afin de détecter des sources d'intérêt pour sa représentation interne de l'environnement. A partir de la définition d'un environnement donnée ci-dessus, la motivation par la réduction de l'Incertitude est celle sur laquelle le modèle HTM se base : nous considérons que la représentation interne de l'environnement que le robot cherche à se construire doit impliquer bonne connaissance des événements/objets audiovisuels pouvant s'y dérouler ainsi que la réduction au maximum de l'incertitude. Mais nous proposons ici une adaptation des paradigmes d'exploration basés sur la *Motivation* en cela que le modèle HTM ne cherche pas à explorer la *topologie* de l'environnement mais son *contenu sémantique*, c'est-à-dire les événements/objets audiovisuels qui peuvent y survenir.

Une autre limite que le modèle HTM tente de dépasser est celle de la façon dont le robot explore l'environnement. Les robots ou systèmes simulés dans lesquels sont implémentés les algorithmes d'exploration prennent en compte un agent mobile dont l'incarnation entière doit se déplacer dans une zone précise. Or ce mouvement est généralement lent, du fait du matériel utilisé ou des algorithmes de détection et d'évitement d'obstacles ou de planification de chemins, et mono-tâche. Nous avons donc tiré parti des mouvements de tête dont le robot que nous avons utilisé est capable, afin d'enrichir l'exploration de l'environnement. En effet, ces mouvements de tête permettent d'effectuer une double tâche d'exploration : une, conduite par le mouvement entier du robot vers une zone d'intérêt, l'autre, conduite par la tête seulement, capable de scanner l'environnement immédiat grâce à ses capteurs visuels pendant que le robot bouge vers une zone d'intérêt. Ces principes de mouvements de tête et de zone d'intérêt font appel aux notions de perception et d'attention.

Dans un premier temps, la section suivante introduit la notion *Perception*, tandis que les phénomènes attentionnels seront traités à la section d'après.

2.2 Perception

La Perception est un ensemble vaste et complexe de phénomènes ayant pour origine la capacité d'acquérir des informations du monde extérieur grâce à nos capteurs anatomiques. Selon le CNRTL [8], la *Perception* est définie comme suit :

1. *Opération psychologique complexe par laquelle l'esprit, en organisant les données sensorielles, se forme une représentation des objets extérieurs et prend connaissance du réel,*
2. *Ce qui est perçu par l'intermédiaire des sens.*

De cette définition lexicale, nous tirons la base primordiale à partir desquels les phénomènes perceptifs émergent : les *sens*. Ces *sens* (audition, vue, toucher, goût, odorat, mais aussi proprioception, sens de l'équilibre, thermoception etc.) permettent la réception de stimuli puis leur transduction en influx nerveux transmis aux aires corticales dédiées à leur analyse. Cette information codée sous forme d'influx nerveux sera ensuite analysée afin de faire émerger des notions cognitives comme celle d'*objet*. D'autre part, cette perception n'est pas unimodale : l'ensemble des informations collectées sont regroupées afin de faire émerger la notion d'objet multimodal ou de concept. Cette pluralité dans la définition des éléments qui constituent le monde extérieur permet de les définir très spécifiquement mais également de pouvoir reconstruire une représentation globale d'un objet, par exemple, à partir d'une partie de ses caractéristiques. Par exemple, le fait de *voir* un objet provoque l'émergence d'un grand nombre d'informations collectées à propos de ce type d'objets ou de cet objet en particulier : le son produit lorsqu'on le tape, la manière qu'il aura de se casser s'il tombe, son poids, sa texture, ou encore des événements en lien avec cet objet.

Du côté de la communauté robotique et de son histoire, il faut remonter aux années 40 pour trouver les premières traces de travaux de recherche sur la robotique autonome disposant de perception. En 1948, le neurophysiologiste William Grey [68] élabore deux robots « tortues », Elsie & Elmer, capables de bouger en fonction de stimuli lumineux ou sonores. L'idée est de reproduire une forme simple de réflexe conditionné. La création de tortues électroniques a ensuite encouragé de nombreux chercheurs à développer ce qui est appelé la *vie artificielle*. Leur capacité à explorer leur environnement (d'où leur surnom qui leur a été parfois donné de *Speculatrix Machina* [69]), leurs réactions de type « réflexe » face à des stimluli lumineux variant en intensité, ainsi que leur comportement adaptatif, ont en quelque sorte marqués l'avènement de la robotique et des robots. Un peu plus tard, en 1986, Valentino Braitenberg, dans son livre *Vehicles : Experiments in Synthetic Psychology* [9] [70], décrit une série d'expériences au cours desquelles des robots mobiles extrêmement simples sont capables d'adopter un comportement complexe grâce à des capteurs

8. Centre National de Ressources Textuelles et Lexicales, créé par le CNRS, est un ensemble de ressources linguistiques informatisées et d'outils de traitement de la langue, cf. `http://www.cnrtl.fr`

9. Véhicules : Expériences en Psychologie de Synthèse.

connectés aux parties motrices par des connexions similaires à des réseaux de neurones. Les véhicules de Braitenberg sont devenus le premier exemple de méthodes réactives : une paire de capteurs visuels est directement connectée à une paire de roues faisant émerger un comportement de type *réflexe*, où la perception est directement associée à l'action. Une étape importante a été franchie par Rodney Brooks lorsqu'il crée le concept de *robotique réactive*. Selon cette nouvelle approche, la perception devient alors le problème central là où il n'était alors considéré que comme secondaire. L'intérêt de la robotique réactive réside également dans la volonté de limiter la nécessité d'apprendre aux robots des *modèles* du monde, restreignant ainsi drastiquement leur adaptabilité à de nouveaux environnements. C'est ici que naît l'approche *comportement-centrée*[10] [71].

Dans le cadre du modèle HTM, nous définissons justement l'environnement comme l'ensemble des objets audiovisuels qui le composent, objets qui causeront une réaction comportementale du robot, concrétisée par un mouvement de tête. Cette section est donc dédiée à la description des phénomènes liés à la perception audio et visuelle ainsi qu'à la présentation de théories sur la façon dont les informations perçues sont traitées, combinées et analysées.

La **Sec. 2.2.1** et la **Sec. 2.2.2** sont dédiées à la description des modalités sensorielles audio et visuelles, chez l'homme notamment, ainsi que leur traitement par les aires du cerveau correspondantes.

La **Sec. 2.2.1.3** est dédiée à un état de l'art sur le rôle des mouvements de tête sur l'audition, notamment dans la localisation de sources sonores.

La **Sec. 2.2.3** présente la Théorie de la Hiérarchie Inverse, théorie innovante établissant un lien direct entre la complexité d'une information à analyser et la réquisition descendante des structures d'analyses nécessaire à la compréhension de cette information.

La **Sec. 2.2.4** introduit la façon dont les informations issues de nos différents capteurs — yeux et oreilles principalement — sont intégrées dans des structures cérébrales dédiées afin de faire émerger la notion d'entité multimodale et même de générer une réaction motrice consécutive.

2.2.1 Audition

L'oreille est une structure anatomique extraordinaire permettant de capter des ondes acoustiques complexes, de les amplifier puis de les transduire en information traitable par le cerveau afin d'en tirer le sens qu'elles contiennent. Pour bien comprendre les défis que l'audition humaine — et animale en général — parvient à relever avec brio, la comparaison avec le système visuel est intéressante. Dans une image, ou une séquence d'images, la décomposition sémantique en *entités* qui composent cette image est physique. Des notions comme la singularité locale, les différences de contraste, l'unité sémantique, permettent de distinguer des flux d'informations visuelles de façon aisée. De plus, les données de localisation spatiale étant perçus directement par les cellules de la rétine, il est possible très rapidement et aisément d'effectuer une ségrégation des entités présentes dans la scène visuelle. A l'opposé, un son, ou une suite de sons, contient toutes les entités présentes en même

10. « *Behavior-based approach.* »

temps ainsi que leur caractéristiques respectives : identité et localisation spatiale notamment. Ainsi, la ségrégation de ces entités ne peut pas se faire au niveau des capteurs, comme l'œil est capable de faire très rapidement, mais doit se faire après un nombre important de transformations et d'analyses effectuées après transmission des informations auditives dans les aires sensorielles dédiées. D'autre part, le son est intrinsèquement une donnée temporelle : là où l'analyse d'une image fixe apporte un grand nombre d'informations, l'analyse du son implique obligatoirement une intégration temporelle et donc la capacité de suivre les différents sons constituant la scène audio. Parvenir à détecter plusieurs sources sonores en même temps, leur attribuer une position spatiale ainsi qu'une identité est un défi que l'oreille humaine — et ses structures analytiques cérébrales dédiées — parvient à relever avec d'incroyables performances en comparaison des systèmes artificiels et des plateformes robotiques binauraux actuels, tant du point de vue de la bonne reconnaissance des sons que de la rapidité avec laquelle ces informations sont traitées.

Cette section est ainsi dédiée à la description de l'audition, en tant que modalité sensorielle et processus perceptif : description anatomique, traitement des informations acoustiques par les aires sensorielles, indices binauraux permettant la localisation des sources sonores, structures cérébrales impliquées dans la reconnaissance.

2.2.1.1 Système auditif humain

Le système auditif humain est composé d'une partie périphérique, le « capteur », et d'une partie centrale, le « centre d'analyse ». La première est dédiée à la capture de l'information audio mais effectue également une première analyse des signaux perçus. La seconde partie est dédiée à l'analyse complète des informations envoyées par la partie périphérique et participe également à son intégration avec les autres modalités sensorielles afin d'aboutir à une représentation plus complexe et riche de l'information audio perçue.

Partie périphérique La partie périphérique, l'« oreille », du point de vue anatomique, est composée de :

l'oreille externe : constituée du pavillon et du conduit auditif externe, l'oreille externe est bien plus qu'une simple structure permettant de faire converger le son vers l'oreille moyenne. En effet, elle est impliquée dans de nombreuses transformations du son lui parvenant : atténuations, amplifications, diffractions et réverbérations ; transformations qui sont fonctions de la position relative de la source sonore et de son contenu fréquentiel. De plus, depuis, les travaux de William D. Batteau en 1967 [72] notamment, son rôle dans les capacités de localisation du son a été amplement étudié et observé : position avant-arrière, localisation en azimut et information sur l'élévation [11]. Le conduit auditif externe, quant à lui, permet d'augmenter spécifiquement le gain des fréquences principales des sons de paroles (entre 2 kHz et 5, 5 kHz). Enfin, les variations de pression ainsi transformées et amplifiées arrivent au

11. bien que les performances de l'humain en estimation de l'élévation d'une source soient relativement mauvaises.

tympan et le font vibrer.

l'oreille moyenne : constituée du tympan, de la chaîne des osselets marteau / enclume / étrier, des cavités mastoïdiennes et de la trompe d'Eustache. Le rôle — et l'intérêt majeur — de l'oreille moyenne est d'adapter la faible impédance de l'air (conduit auditif externe) à la forte impédance de l'endolymphe, liquide constituant la cochlée (partie de l'oreille interne). La chaîne des osselets se termine à la membrane de la fenêtre ovale à qui elle transmet les vibrations perçues par le tympan.

l'oreille interne : constituée principalement du vestibule et des trois canaux semi-circulaires (partie du système vestibulaire responsable de l'équilibre) et de la cochlée. Cette dernière, de structure hélicoïdale remplie du liquide endolymphatique, est celle qui nous intéresse en cela que son architecture est responsable de la transduction des mouvements de l'endolymphe en influx nerveux. En effet, l'intérieur de la cochlée est tapissé des cellules ciliées internes dont la position dans la cochlée les rend sensibles à une bande fréquentielle précise. Ainsi, lorsque l'endolymphe est mis en mouvement par les vibrations du tympan causant ceux de la membrane de la fenêtre ovale, les cellules ciliées sont également mises en mouvement. Ce mouvement ciliaire a pour conséquence une décharge électrique nerveuse rendant compte de cette bande fréquentielle à laquelle les cellules ciliées ont été activées. L'ensemble des nerfs liés à chacune des cellules ciliées internes se regroupent au sein du nerf auditif qui projette directement sur le cortex auditif primaire. La cochlée a une organisation *tonotopique* en cela qu'il y a un lien entre sensibilité d'une cellule ciliée interne à une fréquence et la position de cette cellule dans la cochlée. Nous retrouverons cette tonotopie jusque dans les aires centrales (cortex auditif).

La partie périphérique permet ainsi de capter des ondes acoustiques, de les amplifier et de les transduire en un code neuronal interprétable pour la partie centrale. La structure de l'oreille externe permet également déjà d'extraire quelques caractéristiques sur la position de ou des sources sonores.

Nous sommes ici à la frontière entre le système périphérique et le système central de l'audition. D'une onde acoustique complexe, nous parvenons à sa décomposition en ses fréquences constitutives et en ses caractéristiques temporelles. Cette information est convoyée vers le cortex auditif, dont l'architecture et le fonctionnement sont détaillés au paragraphe suivant.

Partie centrale La compréhension que la perception auditive était traitée par des aires cérébrales remonte à Thomas Willis en 1664 [73] puis en 1681 [74] :

> « *The impression of the sound or the Species admitted to the Ears... [is] carried inwardly towards the Cerebel and* sensorium commune [12]*. [...] Ideas of sounds conveyed also to the Cerebel; which forming there footsteps or tracts, impress a remembrance of themselves, from whence when*

12. Cervelet ainsi que d'autres structures cérébrales plus haut-niveau comme le striatum, placé sous le cortex

afterwards the Species there laid up are drawn forth by the help of the vocal process, voices, like the sounds before admitted, and breaking forth in a certain ordained series, come to be made. »

Cette intuition que Thomas Willis a eue le place comme un des précurseurs de la compréhension des mécanismes cérébraux de l'analyse des stimuli audio. Il faut attendre ensuite la fin du XIXè siècle pour parvenir à une localisation précise du cortex auditif, grâce aux travaux anatomiques sur des singes, de Sir David Ferrier [75] : cette aire cérébrale est située, dans chaque hémisphère, au niveau du lobe temporal. Le cortex auditif est constitué du cortex auditif primaire, première étape de l'analyse du son : hauteur fréquentielle, intensité, durée et timbre ; et du cortex auditif secondaire réalisant une analyse plus poussée : analyse des sons de parole, mémoire sémantique, musique, reconnaissance etc. Une des principales hypothèses sur la façon dont les aires corticales auditives sont organisées est celle d'une organisation en deux voies : le système « *What* » & « *Where* ». Le cortex frontal inférieur est dédié à l'identification de sons (« *What* ») tandis que les aires pariétales sont dédiées à la localisation du son [76] (« *Where* »). Nous reviendrons plus tard un peu plus précisément sur la façon dont le système auditif (et visuel) traite les données audio. Pour lors, la section suivante introduit la notion d'indices binauraux, informations d'importance servant de base dans l'analyse de toute scène acoustique que ce soit par le cerveau ou les systèmes artificiels.

2.2.1.2 Indices binauraux

Les indices binauraux sont toutes les caractéristiques du signal extraites grâce à la conjugaison des informations recueillies par les deux oreilles. Les deux principaux indices binauraux sont la *Interaural Time Difference*[13] et la *Interaural Loudness Difference*[14]. L'ITD est la différence entre le temps que met un son à être perçu par une oreille et ce même son par l'autre oreille. L'ILD est la différence d'intensité du son perçu par chacune des oreilles. Ces différences sont exploitées afin d'estimer la localisation d'une source sonore : une source sonore placée sur la gauche sera perçue comme plus forte par l'oreille gauche que par l'oreille droite (jusqu'à 35 dB de différence, selon [77]) et le son parviendra à l'oreille gauche avant l'oreille droite. Ces indices binauraux sont également fonction de la hauteur fréquentielle des sons perçus. Jens Blauert [78] a montré, sur la base de la théorie Duplex proposée par Lord Raylegh en 1907 [79], que l'ITD est plus robuste et pertinent dans les basses fréquences et, à l'inverse, l'ILD, pour les hautes fréquences. En effet, les ondes basses fréquences sont facilement convoyées autour de la tête produisant des valeurs d'ILD négligeables (cf. **Fig. 2.12**), tandis qu'elles provoquent des ITD produisant des différences interaurales de phase détectées par les neurones du tronc cérébral. En revanche, pour les signaux à large bande fréquentielle, il semble que l'ITD soit malgré tout dominant et que la fréquence à laquelle une séparation est possible entre la dominance de l'utilisation de l'ITD ou de l'ILD soit située aux environs de 2 kHz (variable selon les études et les auteurs). D'autre part, en 2000, Barbara Shinn-Cunningham *et al.* [80] ont montré que l'ILD est particulièrement informatif pour

13. ITD — Différence interaurale de temps.
14. ILD — Différence interaurale d'intensité ou différence interaurale de phase.

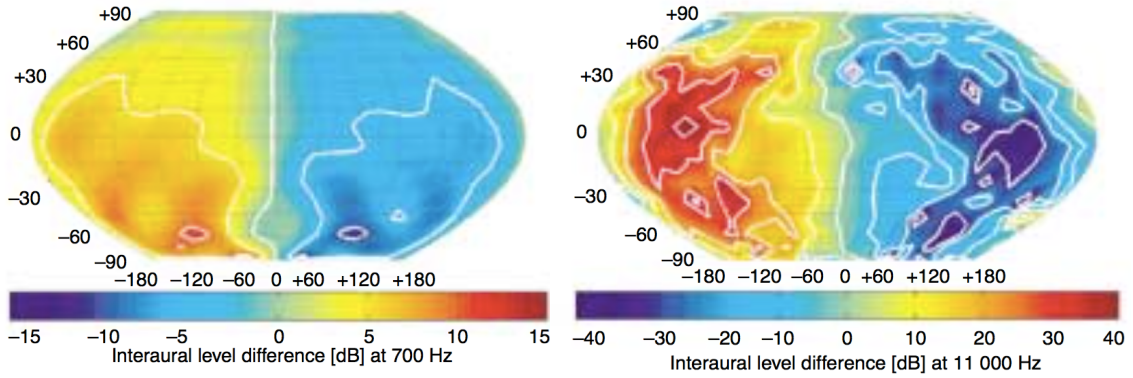


FIGURE 2.12 – DIFFÉRENCES INTERAURALES D'INTENSITÉ — Valeurs d'ILD (« *Interaural Loudness Difference* ») mesurées chez un sujet humain en fonction de la direction d'une source sonore, (*gauche*) pour un son de basse fréquence (700 Hz) et (*droite*) pour un son de haute fréquence (11 kHz) mettant en avant la différence d'informativité de l'ILD en fonction de la fréquence (figure d'après [81]).

---

les sources sonores éloignées d'une distance comprise entre un et deux mètres de la tête et à proximité de l'axe interaural.

SHINN-CUNNIGHAM *et al.* ont également étudié intensivement le *cône de confusion*, zone conique de l'espace dont l'origine est le point situé entre les deux oreilles et qui rayonne de chaque côté, à l'intérieur duquel les indices binauraux sont très peu informatifs pour la localisation d'une source sonore. Une source sonore placée à cet endroit aboutit à des ambiguïtés perceptuelles rendant la localisation difficile. Malgré tout, même placées dans cette zone particulière, l'humain est capable de localiser des sources sonores avec performance : mais l'audition seule n'est alors plus suffisante. C'est ici que les mouvements de tête apparaissent, augmentant considérablement les capacités de localisation du son dans des contextes difficiles ou dans des situations acoustiques particulières. L'impact des mouvements de tête sur la perception auditive est décrit à la section prochaine.

2.2.1.3 Mouvements de tête et Audition

De quelle manière les mouvements de tête peuvent-ils être bénéfiques à l'audition ? Selon JENS BLAUERT [78], les mouvements du corps, et en particulier les mouvements de tête, participent à la performance des animaux dans la localisation de sources sonores. Cette section détaille les travaux de recherche ayant mis en avant l'intérêt de ces mouvements de tête dans l'amélioration de l'écoute et de l'analyse conséquente des signaux audio perçus.

Au début des années 1930, PAUL THOMAS YOUNG a été un des premiers à étudier le rôle et l'impact des mouvements de tête dans la perception du son. En 1931 [82] YOUNG a mené une série d'expériences à l'aide d'un dispositif matériel fixé sur la tête de sujets volontaires simulant le cône de confusion dans le but de supprimer l'impact des mouvements de la tête et du corps sur le calcul des indices binauraux permettant la localisation du son. La position des sources apparentes perçue par les sujets a montré une imprécision générale dans la localisation, excepté pour la discrimination

gauche-droite. Une importante confusion avant-arrière a été observée ainsi qu'une mauvaise perception de l'élévation. Cette expérience a été une des premières mettant en avant le lien entre mouvements de tête et localisation spatiale de sons : le dispositif incapacitant les sujets humains a résulté en la diminution de la capacité à utiliser les indices binauraux qui auraient pu être apportés par des mouvements de tête.

En 1938, Hans Wallach [83] a également mené des expériences dans ce domaine visant à étudier l'apport des mouvements de tête dans l'extraction d'informations provenant du cône de confusion. Wallach a défini la distance angulaire entre la source sonore et l'axe des oreilles comme l'*angle latéral* et émit l'hypothèse selon laquelle les mouvements de tête, par modification de cet angle latéral, permettaient de lever les ambiguïtés causées par le cône de confusion, la confusion avant-arrière ou l'élévation de la source sonore.

En 1967, Williard R. Thurlow *et al.* [84] conduisit des expériences subjectives afin d'étudier spécifiquement les différents mouvements de tête en fonction de sons ne contenant que des fréquences basses ou que des fréquences hautes. Les mouvements de tête considérés ont été la rotation axiale, l'inclinaison et la flexion/extension. Les résultats de Thurlow *et al.* montrent que (i) la rotation est le mouvement le plus fréquent des mouvements uniques, parmi les trois, (ii) la rotation combinée à la flexion/extension était le mouvement le plus fréquent, (iii) la portée des mouvements était la plus forte pour les mouvements de rotation, et (iv) la rotation maximale a été observée majoritairement pour les signaux basses fréquences, appuyant par ailleurs le fait que le contenu fréquentiel a une influence sur la capacité de localisation. D'autre part, la plupart des mouvements observés avaient pour direction la source sonore à localiser. En parallèle de cette étude, Williard R. Thurlow & Philip S. Runge [85], la même année, ont étudié l'apport des mouvements de tête dans la précision de la localisation d'une source sonore. Trois conditions ont été testées : une pour laquelle quatre types de mouvements de tête étaient autorisés (*via* une contrainte matérielle), une deuxième pour laquelle la tête était libre de bouger librement, une dernière pour laquelle aucun mouvement n'était possible. Les quatre mouvements de tête sont : la rotation, l'inclinaison, la combinaison rotation/inclinaison, et la flexion/extension. Les sons utilisés sont du bruit basse ou haute fréquence et de durées variables. Les résultats des expériences conduites montrent une amélioration significative de la localisation horizontale lors des rotations, indifféremment du type de son. En revanche, la diminution des erreurs de localisation sur le plan vertical était très faible. Enfin, la condition en mouvement libre a résulté en une amélioration mineure de la localisation, même comparée à la rotation seule.

En 1997, Stephen Perrett & William Noble [86, 87] ont montré, selon un paradigme expérimental très proche de ceux employés par Thurlow, que la confusion avant-arrière, prévalente lorsque les mouvements de tête étaient impossibles, est quasiment éliminée par la rotation de la tête. Utilisant des sons au contenu fréquentiel un peu différent, les auteurs ont montré que les fréquences inférieures à 2 kHz étaient nécessaires pour que les rotations de la tête soient efficaces en localisation verticale. En addition à ce résultat, ils ont observé que l'effet des rotations était plus fort pour les sources placées dans le plan vertical face au sujet. L'ensemble de ces résultats, par ailleurs, concordent et appuient les hypothèses et observations de Wallach, de Thurlow et de Runge sur le rôle des mouvements de rotation dans

les capacités de localisation.

En 2013, Chungeun Kim *et al.* [88] ont conduit des expériences similaires mais en testant des volontaires sur la largeur de la source sonore, l'enveloppement et le timbre, en plus de la localisation de sources sonores. Durant les expériences, les mouvements de tête ont été suivis et enregistrés. De façon étonnante, les auteurs ont montré que les sujets ont effectué plus de rotations pour juger la largeur et l'enveloppement que pour la direction ou le timbre. De plus, les sujets ont tendance à essayer de faire face aux sources sonores que ce soit pour la localisation ou pour la largeur et l'enveloppement des sons. D'autre part, des expériences additionnelles ont été menées dans des environnements acoustiques plus réalistes comme écouter un concert, jouer à des jeux vidéos ou regarder des films. Les résultats obtenus sont similaires à ceux présentés précédemment.

2.2.1.4 Systèmes auditifs artificiels

La création d'un système auditif artificiel a généralement pour but de donner à un robot la capacité de percevoir et de traiter les signaux sonores. Selon Richard F. Lyon dans sa revue publiée en 2010 [89] un système auditif artificiel doit comprendre :

une analyse périphérique : consistant en l'acquisition du son (oreille externe), en un traitement fréquentiel (conduit auditif externe, chaîne des osselets, cochlée) et en sa transduction en signal interprétable (cellules ciliées internes) ;

l'extraction de caractéristiques : consistant en la détermination d'indices binauraux tels que l'ITD ou l'ILD ;

l'interprétation : permettant de donner un sens aux analyses précédemment effectuées, comme par exemple estimer la position d'une source à partir des indices binauraux de localisation ;

la cognition : consistant en l'intégration de différentes sources d'informations couplées au résultat de l'étape d'interprétation et permettant de faire émerger un comportement, une réaction ou une prise de décision consécutive et causée par le contenu sémantique du son perçu.

Deux approches ont alors émergé reposant sur des conceptions opposées de la robotique ainsi que des ambitions différentes. La première est celle de doter une plateforme robotique d'une antenne de microphones (c'est-à-dire de plus de deux microphones) disposés selon des géométries variables et permettant, par multiplication de la redondance des informations sonores captées, d'augmenter significativement les capacités d'analyse audio. Cette approche a pour but de permettre au robot d'être le meilleur possible pour l'analyse de scènes sonores. La seconde, à l'opposée, est celle de s'inspirer du système auditif des mammifères, et de l'humain en particulier, se basant sur le fait que si l'évolution a convergé (pour le moment) vers une audition binaurale, il doit être possible de créer des systèmes dotés de seulement deux microphones aux mêmes performances de traitement du son que l'humain.

Mais au-delà de la différence de philosophie adoptée lorsqu'il est question de développement de systèmes perceptifs et/ou intelligents, des contraintes techniques sont aussi à prendre en compte lors de l'intégration de tels systèmes au sein d'une pla-

teforme robotique. Par exemple, en 2013, Sylvain Argentieri *et al.* [90] pointent celles-ci :

embarquabilité : les capteurs artificiels (microphones ou pavillons manufacturés) doivent pouvoir être intégrés à la plateforme robotique sans mettre en péril les capacités du robot (mouvements de tête, équilibre du robot, mobilité. . .). L'approche binaurale permet de remplir cette condition tandis qu'à l'opposé, les antennes de microphones prennent une place significative, leur efficacité étant proportionnelle à leur nombre.

temps réel : la puissance du système auditif humain repose en partie sur la rapidité de l'analyse des signaux sonores. Mais plus qu'une performance, il s'agit d'un besoin écologique pour pouvoir réagir avec pertinence aux événements sonores perçus [91]. Une nouvelle fois, l'approche binaurale présente un intérêt en cela que le temps computationnel nécessaire afin d'analyser les signaux sonores perçus est moindre que pour une antenne de microphones (même si cette différence tend à se réduire de plus en plus).

morphologie : la capacité de l'humain a utiliser des mouvements pour améliorer l'analyse d'une scène audio est majeure. La création d'un système artificiel doit pouvoir également intégrer cet aspect actif de la perception auditive. Ici aussi, la plus grande compaction des systèmes binauraux permet de les intégrer à des plateformes mobiles mimant des mouvements observés chez l'homme ou l'animal.

environnement : une scène sonore est éminement complexe, notamment par la présence de bruit de fond, de réverbération ainsi que par l'éventuelle mobilitédes sources sonores. De plus, le robot lui-même, du fait de sa motorisation, génère du bruit venant également parasiter la scène audio. Un système auditif artificiel doit être capable de filtrer ces bruits auto-générés ainsi qu'être robuste à l'ensemble des éléments perturbateurs de la scène audio.

Les deuxième et troisième points sont ceux qui nous intéressent particulièrement. Premièrement, le modèle HTM, en tant que système capable de déclencher des mouvements de tête, dote le robot d'une capacité de perception active d'un environnement acoustique et participe ainsi à l'amélioration de l'analyse de cette scène. D'autre part, la notion de *temps réel* est centrale dans le cadre du modèle HTM, qu'elle soit comprise dans le sens très strict des roboticiens (prenant en compte le décours temporel des actions motrices à effectuer, une latence dans la transmission des données ou le temps computationnel pour l'analyse des informations — toutes ces durées éloignant le moment où l'information est captée du moment où un sens utile en est extrait) ou dans le sens plus variable des neurosciences (où la notion de temps réel est comprise comme « un temps très court »). En effet, une de ses ambitions — et un de ses intérêts — est de pouvoir réagir extrêmement rapidement aux événements audio, visuels ou audiovisuels présents dans un environnement inconnu. Ainsi, parvenir à extraire les informations nécessaires à la réaction du robot suffisamment rapidement est indispensable. Le paradigme utilisant des antennes de microphones ne sera pas décrit ici plus en détail puisque le robot dont nous disposons est doté d'une tête binaurale.

2.2.2 Vision

A la différence du système auditif externe, la vision est dotée d'un système externe — les yeux — effectuant une première analyse de la scène visuelle. En effet, les cellules de la rétine permettent déjà d'extraire de nombreuses caractéristiques de la scène telles que l'intensité, le contraste, le mouvement ou la couleur. Cette primo-analyse poussée est également rendue possible par la façon dont les informations visuelles sont organisées, dans le monde physique. Premièrement, l'analyse d'une scène visuelle n'a pas forcément besoin d'être intégrée temporellement : des entités sont déjà discernables dans une image fixe, contrairement à l'audio, et l'ensemble d'une scène complexe peut être analysée avec cette seule image. Il est même possible de percevoir une notion de mouvement potentiel avec une seule image. Cette différence est majeure. Deuxièmement, les entités présentes sont souvent beaucoup moins interférantes que dans une scène audio : deux objets distincts dans une image ne se superposent pas, à la différence de deux objets audio. Et dans le cas où elles seraient interférantes, gênant ainsi leur discrimination, le système visuel humain, couplé aux mouvements de tête, est capable de réduire son champ d'analyse de la scène drastiquement afin de se concentrer sur une partie seulement de l'espace visuel : modification du point de focalisation, mouvements des yeux, modication de l'ouverture de l'iris, notamment. Toutes ces différences confèrent au système visuel une précision de l'analyse d'un environnement largement supérieure aux capacités du système auditif, tant en qualité d'analyse qu'en rapidité ou en robustesse.

La **Sec. 2.2.2.1** est dédiée à la description succinte du système visuel humain (parties périphérique et centrale). La **Sec. 2.2.2.2** sera quant à elle dédiée à une description des contraintes auxquelles les systèmes de vision artificiels sont soumis.

2.2.2.1 Système visuel humain

Tout comme le système auditif, le système visuel est constitué d'une partie périphérique et d'une partie centrale. Cette section détaille ainsi ces deux parties dont l'architecture et l'organisation sont très proches de celles du système auditif.

Partie périphérique La partie périphérique de la vision humaine commence par le globe oculaire, principalement composé de la cornée, du cristallin et de la rétine. La cornée, modélisée comme une lentille convergente, permet de transmettre la lumière au cristallin et à la rétine. Le cristallin, deuxième lentille convergente biconvexe, permet de focaliser les rayons lumineux sur la rétine, s'adaptant à la distance de l'objet visuel perçu. La rétine, enfin, est la partie réceptrice des stimuli lumineux, l'équivalent d'un écran sur lequel la lumière se projette. La rétine est composée des cônes, cellules sensorielles responsables de la vision en couleur et de la vision diurne, et des bâtonnets, cellules sensorielles responsables, à l'opposé des cônes, de la vision en noir et blanc et de la vision nocturne. Les cinq millions de cônes et cent-vingts millions de bâtonnets transduisent les stimuli lumineux en signaux électrochimiques captés par des neurones dédiés (les cellules bipolaires) connectés à leur tour aux cellules ganglionnaires se regroupant au sein du nerf optique, porteur de l'information désormais exprimée sous forme d'impulsions électriques.

La partie périphérique de la vision est très similaire à la partie périphérique du système auditif : il s'agit de concentrer les informations à un endroit particulier de l'organe récepteur, de les organiser de façon optimale et de les traduire en code neuronal afin de les transmettre le plus efficacement et le plus rapidement au centre d'analyse. En revanche, une grande différence entre les systèmes visuel et audio est que l'œil est doté de mouvements, grâce aux six muscles oculomoteurs responsables des différents mouvements oculaires comme les saccades (mouvements très rapides) ou la dérive (mouvements lents). Ces capacités motrices sont indispensables pour une perception visuelle correcte pour deux raisons : (i) la vision n'a accès qu'à une partie limitée de l'espace visuel et (ii) la vision n'est précise qu'au niveau de la fovéa, petite partie située au centre de la rétine. Le principal rôle des mouvements oculaires est donc de pouvoir scanner l'espace visuel rapidement et de placer la fovéa au niveau des stimuli visuels d'intérêt afin de les analyser avec précision.

Partie centrale Le nerf optique, à partir de l'œil, atteint le cortex visuel primaire (V1) après être passé par le noyau thalamique géniculé (LGN [15]). 90% des informations visuelles captées par la rétine parviennent au LGN, les 10% restant sont dirigés vers le colliculus supérieur (CS, structure sur laquelle nous reviendrons plusieurs fois dans les sections dédiées à la perception d'une part, et à l'attention d'autre part). Le cortex visuel est situé au niveau du lobe occipital (arrière du crâne) et est constitué du cortex visuel primaire (V1), du cortex strié et d'aires extrastriées (V2-V4, cortex inférotemporal, aire médiotemporale et cortex pariétal postérieur). De nombreuses questions sont encore sans réponses quant à l'organisation et le fonctionnement du cortex visuel primaire [92] et plus encore concernant les aires extrastriées. Une des découvertes les plus importantes a été celle de la nature profondément parallèle de l'analyse des informations visuelles, en opposition à une analyse sérielle. De façon similaire à la tonotopie du cortex auditif, le cortex visuel primaire présente une rétinotopie qui, à deux points proches de la rétine fait correspondre deux points proches de V1. Une quinzaine de cartes rétinotopiques ont été caractérisées chez l'homme, présentes dans les différentes aires du cortex visuel. Chaque carte représente une caractéristique de la scène visuelle : contraste, mouvement, couleur, orientation... également dédiée à un hémichamp visuel particulier.

Plusieurs hypothèses étayées par des résultats expérimentaux ont été émises quant à l'architecture des aires visuelles plus haut niveau impliquées dans le traitement de l'information visuelle. De façon similaire au système auditif, une des hypothèses les plus répandues et émise par Mortimer Mishkin *et al.* en 1983 [93], est celle de l'organisation en deux voies distinctes : l'une traitant le *Quoi ?* (« *What ?* »), c'est-à-dire formes et couleurs principalement, l'autre traitant le *Où ?* (« *Where ?* »), c'est-à-dire position et mouvement, principalement. La voie occipitotemporale, ou « flux ventral », est dédiée à l'identification des objets (« *What* »), tandis que la voie occipitopariétale, ou « flux dorsal », est dédiée à la relation spatiale entre les objets (« *Where* ») ainsi qu'à l'analyse des mouvements visuels vers ces objets [94]. Comme pour la partie centrale du système auditif, la section dédiée à la perception reviendra sur ce point.

15. Laterate Geniculate Nucleus.

2.2.2.2 Systèmes visuels artificiels

La conception d'un système visuel artificiel (SVA) est soumis aux mêmes contraintes que celles exposées pour l'audition à la **Sec. 2.2.1.4**. Les principes, listés par Richard F. Lyon [89], d'analyses périphériques, d'extraction de caractéristiques, d'interprétation et de cognition sont applicables de la même façon pour les SVA.

D'autre part, concernant l'aspect morphologique d'un SVA, plusieurs systèmes existent : caméras ou ensemble de caméras permettant une vision à 360° ou caméra unique au champ de vision restreint ou encore vision binoculaire similaire à la vision des mammifères. De plus, différentes plateformes robotiques existent intégrant la possibilité de faire bouger les yeux. La plateforme robotique utilisée au sein de Two!Ears disposera d'une vision binoculaire n'incluant pas de dispositifs moteurs pour faire bouger les yeux. Ceux-ci étant fixés sur la tête du robot, le mouvement des yeux sera celui de la tête.

Enfin, nous notons le développement des caméras événementielles basées sur une rétine artificielle faite de silicone et dont la conception est largement bioinspirée [95, 96]. Cependant, même si ces caméras changent la façon dont les données sont perçues, notamment *via* des principes de pixels agissant comme des cellules de la rétine et envoyant des trains de *spikes*, la dimension temporelle reste une grande différence entre les systèmes visuels et les systèmes audio : un pixel de ce type de caméra n'enverra d'information que lorsqu'un *changement* de la scène visuelle sera perçu [97]. Ainsi, une scène visuelle dans laquelle aucun changement ne survient porte malgré tout une quantité d'information largement supérieure à une scène audio équivalente.

2.2.2.3 Discussion

Cette section a décrit succintement la modalité visuelle chez l'humain, modalité dont est doté la plateforme robotique utilisée au cours du projet Two!Ears. Nous avons notamment pu voir les nombreuses ressemblances entre le système auditif et le système visuel (partie périphérique et partie centrale, organisation des aires sensorielles en *What & Where*, hiérarchie des processus d'analyse etc.) ; mais également leur différence fondamentale : l'aspect temporel. La vision traite en effet des données dont la dynamique temporelle est complètement différente des données audio : une image seule, captée en quelques millisecondes, porte un très grand nombre d'informations (localisation d'objets visuels en azimuth et en profondeur, identité, mouvement éventuel etc.). A l'inverse, une trame audio captée sur la même durée, ne serait que très peu informative. Cette différence est importante et a motivé certains de nos choix sur le plan de l'intégration multimodale des données audio et visuelles.

Le traitement des informations sensorielles, audio ou visuelles, décrit à cette section met en avant la séparation en deux flux traitant d'un côté la localisation de l'événement perçu et son identité. Cette organisation en un flux ventral et dorsal permet notamment de paralléliser l'analyse d'une scène audio ou visuelle. Cependant, de nombreuses travaux de recherche tendent à montrer que la vitesse d'analyse d'un événement audio ou visuel n'est pas indépendant de sa complexité ainsi que du contexte dans lequel il se situe. Dans le cas de l'audio par exemple, un son est d'autant plus dur à localiser et à identifier qu'il se trouve dans un environnement

multisource, bruité ou très réverbérant. D'autre part, il a également été montré que les aires sensorielles sont capables d'émettre des hypothèses sur la probabilité qu'a un événement perceptif de survenir, étant donné le passé à court-terme observé. Ces capacités prédictives, détaillées à la **Sec. 2.3.1.4**, permettent également d'accélérer l'analyse des stimuli audio et visuels.

Ces mécanismes permettent la modulation de l'analyse effectuée par les aires sensorielles. La section suivante introduit ainsi la Théorie de la Hiérarchie Inverse, théorie récente et innovante proposant un modèle intégrant le niveau de complexité de la scène sensorielle dans les processus d'analyses par les aires corticales dédiées.

2.2.3 Théorie de la Hiérarchie Inverse

Les théories sur l'organisation et l'architecture des aires sensorielles, comme le « *What and Where system* » (cf. **Sec. 2.2**), permettent d'expliquer la façon dont les informations visuelles et auditives sont analysées, depuis les aires bas-niveau, traitement des caractéristiques des signaux, jusqu'aux aires haut-niveau, intégration multimodale et implication de processus cognitifs — mémoire, attention endogène, tâches à accomplir etc. En revanche, ces théories ne prennent pas en compte le degré de complexité des données à traiter et l'éventuelle modification des voies cérébrales empruntées lors de ces traitements en fonction de cette complexité. Le but de la théorie décrite dans cette section est justement de proposer un modèle de traitement hiérarchique des données sensorielles selon lequel le degré de complexité des données à traiter est pris en compte dans la détermination des aires et processus cérébraux nécessaires à leur analyse. Cette prise en compte permet d'accélérer le processus d'analyse en shuntant éventuellement certaines étapes. De plus, rappelant le contexte d'exploration d'environnements inconnus dans lequel nous nous situons, les stratégies d'exploration sont également directement fonction du nombre et de la qualité des informations présentes dans l'environnement. En effet, être capable de distinguer différentes sources audio ou visuelles entre elles constitue une étape clef dans l'élaboration de ces stratégies ainsi que dans la rapidité à prendre une décision, motrice par exemple.

En 2004, Merav Ahissar & Shaul Hochstein [98], en 2006, Israël Nelken & Merav Ahissar [99] et enfin en 2008, Mor Nahum *et al.* [101], ont proposé un modèle de traitement global par le cerveau des informations auditives et visuelles impliquant aussi bien une communication ascendante (*bottom-up*) que descendante (*top-down*) entre les capteurs — auditifs et visuels mais aussi les structures bas-niveau placées juste après — et les aires computationnelles — cortex auditifs et visuels. Ce modèle, nommé *Reverse Hierarchy Theory* (RHT, cf. **Fig. 2.13** & **Fig. 2.14**), tente d'expliquer comment ces structures parviennent à analyser le contenu de scènes audiovisuelles ambigües. La RHT se base sur l'assertion suivante du Pr. Shihab Shamma [100] :

> « *A parsing decision is first based on the highest available level of visual representation*[16] ».

16. « La décision d'analyse [des signaux] est tout d'abord basée sur la représentation visuelle disponible la plus haut-niveau »

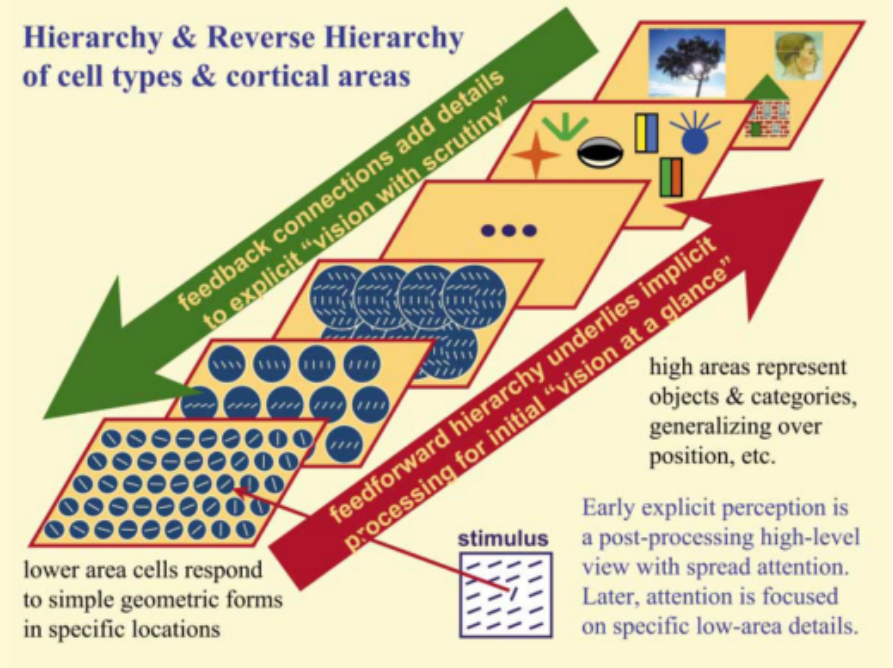


FIGURE 2.13 – THÉORIE DE LA HIÉRARCHIE INVERSE — Illustration de la RHT dans le système visuel, théorie selon laquelle le niveau d'analyse de l'information est directement dépendant de la capacité à analyser cette information (figure d'après [98, 99, 100]).

Dans les systèmes visuel et auditif, cette théorie stipule que la vitesse à laquelle l'information se propage des aires bas-niveau vers les aires haut-niveau dépend de la capacité à discriminer les différentes sources d'information. Dans des tâches complexes de discrimination, seules les caractéristiques bas-niveau des stimuli vont permettre une distinction efficace de deux, ou plus, objets perceptifs concurrents. En revanche, lorsqu'il n'y a pas d'ambiguité, toutes les caractéristiques bas-niveau ne sont pas nécessaires pour analyser et comprendre ces stimuli ; ainsi, l'information se propage-t-elle beaucoup plus rapidement vers les aires haut-niveau. Par exemple, si nous voyons un verre tomber dans un environnement silencieux, le bruit émis lorsqu'il se cassera sera (i) très prédictible et (ii) facilement reconnaissable. Ainsi, selon la RHT, l'analyse de ce son, et de son sens, sera accélérée par l'absence de temps passé à analyser les composantes bas-niveau du signal audio : les calculs d'ILD et d'ITD peuvent par exemple être évités puisque le système visuel a déjà accès à cette information. Mais si le son perçu n'est pas *congru* avec ces prédictions, la RHT postule que ce seront les caractéristiques bas-niveau du signal qui seront nécessaires pour lever l'ambiguité. Et ce comportement n'est possible qu'en la présence de connexions descendantes, allant des aires cérébrales haut-niveau vers les capteurs et leurs structures bas-niveau adjacentes.

Contrairement à de nombreux processus perceptifs traités par le système visuel, les objets audio ne sont pas considérés comme *statiques* mais plutôt comme des flux continus d'information faisant émerger une représentation dynamique de l'objet (voir [100] pour une revue). Chaque flux a sa propre identité perceptuelle. D'après SHIHAB SHAMMA [100] :

> « *The rules and interactions between the stream percepts and the low-*

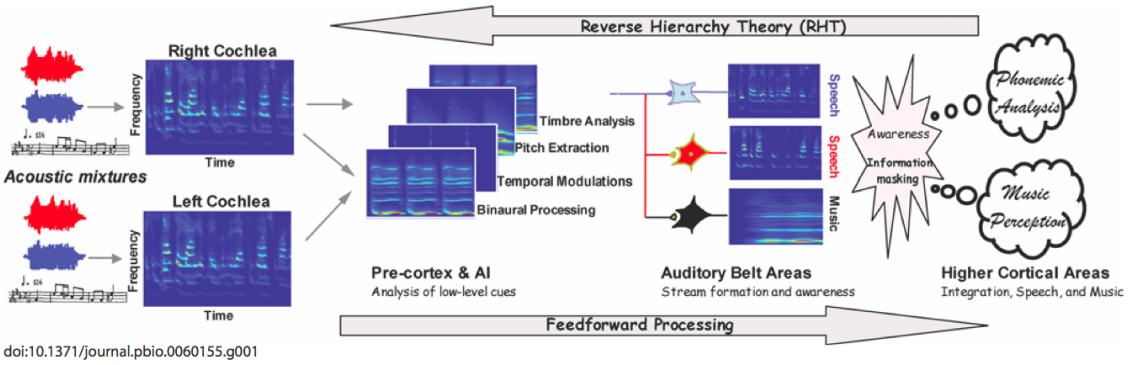


FIGURE 2.14 – THÉORIE DE LA HIÉRARCHIE INVERSE — Illustration de l'adaptation de la RHT dans le système auditif, mettant en avant les flux ascendants et le contrôle descendant de l'information (figure d'après [100]).

> *level cues that group the elements of a stream and distinguish it from its counterparts (e.g., pitch, timbre, and binaural cues) have been delineated over the years under the umbrella of auditory scene analysis*[17]*.* »

La RHT permet ainsi d'analyser les processus attentionnels liés à la perception d'événements sonores et/ou visuels en cela qu'elle relie la nécessité d'allouer de la puissance de calcul à l'ampleur de l'ambiguïté du flux perceptif. Ainsi, il est possible d'accélérer fortement la compréhension de l'environnement tout en diminuant la sensibilité à des stimuli potentiellement non pertinents ou non informatifs. Cela soulève ainsi plusieurs questions à propos de la nature d'un objet :

1. à quelle étape de l'analyse cérébrale un objet est-il reconnu ? Ou, en d'autres termes, quand la notion d'objet émerge-t-elle au sein du flux perceptif ?
2. combien d'indices sensoriels sont nécessaires pour avoir une représentation interne d'un objet suffisamment robuste et stable pour prendre une décision ?

La Théorie de la Hiérarchie Inverse, appliquée aux systèmes visuel et auditif, permet de mettre en avant une propriété importante de la perception : *l'analyse* complète *d'un stimulus n'est requise que si elle est nécessaire.* Cette considération a de grandes conséquences pour la compréhension de la perception visuelle et auditive. En effet, elle implique que le cerveau se base autant que possible sur la représentation interne des objets audio ou visuels qu'il a déjà aquise et en laquelle il fait confiance. Ce n'est que lorsqu'une ambiguïté apparaît que l'analyse profonde, sur la base des caractéristiques bas-niveau des signaux, est requise. Cette ambiguïté se manifeste soit par la complexité de la scène à analyser, par l'incapacité à discriminer le stimulus dans l'environnement perceptif ou encore par l'imprédictibilité de son apparition. La RHT a largement inspiré la conception du modèle HTM en cela qu'elle se base en partie sur la notion de *Congruence* d'un événement au sein d'une séquence prédictible de stimuli ou au sein d'un environnement. La présence d'une incongruence va en effet occasionner la réquisition des données brutes afin que le signal, dont l'analyse avait été accélérée du fait de sa prétendue prédictibilité, puisse être pleinement ana-

17. « Les règles et interactions entre les flux perceptuels et les caractéristiques bas-niveau qui rassemblent les éléments d'un flux et le distinque de ses homologues (par exemple, hauteur, timbre ou indices binauraux) on été délimités au cours des années par l'analyse de la scène audio. »

lysé. Concept fondamental de notre modèle, nous définirons la Congruence comme la prédictilité d'un événement audiovisuel au sein d'un environnement défini par les objets audiovisuels qui y sont présents. La description des principes neuronaux sous-tendant l'existence de cette notion de Congruence seront détaillés à la section dédiée à l'attention, en cela que nous l'utiliserons comme base de la partie attentionnelle du modèle HTM. Mais la définition des événements perceptifs sur laquelle le modèle se base étant principalement audiovisuelle, donc multimodale, il est nécessaire de détailler les mécanismes neuronaux responsables de l'intégration des différentes sources d'information — nos capteurs — et leur combinaison, dans le but de faire émerger une représentation complexe et diverse des événements perceptifs présents dans l'environnement.

2.2.4 Intégration multimodale

L'intégration multimodale, ou cross-modale, est le rassemblement des différentes sources d'information perceptives afin de faire émerger une représentation coginitive plus complexe et plus riche. Nous avons vu plus haut que l'audition et la vision sont des modalités sensorielles qui, malgré leur apparente ressemblance, sont très différentes : fonctionnement des capteurs, intégration temporelle, modulations des mécanismes d'acquisition des informations, capacités des capteurs etc. Beaucoup de modèles computationnels de la perception auditive, en particulier, tentent de créer des systèmes artificiels aux performances extraordinaires : localisation multisources, identification en environnement bruité et/ou réverbérant, analyse de scène audio sans information *a priori* sur l'environnement etc. Souvent, une des justifications est l'observation, chez l'homme, de hautes performances dans l'accomplissement de ces tâches complexes.

Cependant, nous pensons l'inverse. Nous considérons plutôt que le système auditif est une modalité relativement peu précise en comparaison de la vision, non pas du fait de la façon dont la perception auditive est construite, du point de vue cérébral, mais à cause de l'extrême complexité d'une scène audio par rapport à une scène visuelle. D'ailleurs, lorsqu'il s'agit de juger les capacités du système auditif humain et de justifier l'élaboration de systèmes artificiels audio aux ambitions démesurées, la *cécité* et l'adaptation consécutive du système auditif sont très souvent citées. Or, cet exemple n'est, selon nous, pas pertinent, pour deux raisons. La première est que la privation d'une modalité sensorielle — et particulièrement la vision — est une contrainte écologique énorme mettant en péril les capacités de survie d'un individu. Ainsi, cette contrainte aboutit à de nombreux remaniements corticaux [102] afin de palier l'absence majeure de cette source d'information et permettre au sujet aveugle de continuer à avoir une représentation de son environnement la plus correcte et précise possible. Cette adaptation n'est pas un exemple pertinent à prendre puisqu'il est le résultat d'une situation extra-ordinaire. La deuxième raison est que les hautes performances des individus aveugles lors de l'accomplissement de tâches audio n'est pas un phénomène observé dans tout le spectre des tâches auditives. En effet, bien qu'une amélioration significative des capacités de localisation a été observée et mesurée dans nombres d'études [103, 104], il existe également certains cas où l'absence de vision empêche les individus aveugles d'effectuer certaines tâches légèrement plus complexes. En 2014, Monica Gori et al. [105] ont effectué des expériences de bi-

section spatiales (dans une séquence de quatre sons, identifier la localisation perçue des deux derniers sons par rapport aux deux premiers) ou temporelles (identifier si, dans une séquence de trois sons, le deuxième est temporellement plus proche du premier ou du troisième) chez des individus aveugles et voyants. Les performances des individus aveugles pour la tâche de bisection spatiale étaient les mêmes que celles des individus voyants ; pour la tâche de bisection temporelle en revanche, les aveugles ont même montré une incapacité à effectuer la tâche, montrant un déficit probable dans l'implication de la mémoire de travail pour la réalisation d'une telle tâche.

Enfin, un point majeur est qu'un individu voyant et entendant dispose justement de ces deux modalités, il n'est presque jamais nécessaire de mobiliser toutes les capacités unimodales disponibles afin de percevoir correctement l'environnement. Imaginons par exemple que nous marchons dans une rue à sens unique, dans le sens des voitures, la chaussée située à droite. Le fait de (sa)voir que la chaussée est justement sur la droite et que la voie est à sens unique permet d'émettre un certain nombre d'hypothèses sur les éventuels sons pouvant apparaitre : les sons de moteurs de voiture ont (i) une très grande probabilité d'apparaitre, (ii) particulièrement dans le demi-espace droit et (iii) provenant de l'arrière et se dirigeant vers l'avant. Toutes ces hypothèses permettent de simplifier énormément l'analyse de scènes multimodales complexes. En faisant un parallèle avec l'identification d'un son par un système artificiel, résumé et simplifié en une tâche de recherche d'une étiquette sonore au sein d'une très grande base de données d'étiquettes, les hypothèses émises sur l'environnement grâce à la vision permettent de restreindre drastiquement cette base de données.

L'audio, considérée comme une modalité foncièrement différente de la vision, est alors indispensable à l'analyse d'une scène multimodale en cela que le système auditif agit comme un système d'analyse à 360° de la scène permettant de déclencher une réaction à des stimuli d'intérêt présents là où la vision n'a aucune information. Audition et vision travaillent de façon très rapprochée, la vision profitant des capacités d'analyse spatiales de l'audio, et l'audio profitant de la rapidité et de la précision de l'analyse du contenu par la vision (voir également la **Sec. 2.3.1**). Ainsi, l'intégration multimodale est bien plus qu'une mise en commun des informations perçues : elle doit être comprise comme une collaboration très forte permettant d'avoir une représentation de la scène dans l'espace entier.

2.2.4.1 Bases neurales et psychophysiques

Du point de vue cérébral, plusieurs structures ont été identifiées comme recevant des afférences de plusieurs modalités sensorielles : le sillon temporal supérieur, le lobe pariétal ou le putamen par exemple. Mais Alex Meredith & Barry E. Stein, en 1986 [106] affirment, sur la base d'une revue des travaux de recherche sur l'intégration multimodale dans le cerveau des mammifères, que le colliculus supérieur (CS) est une des structures majeures de l'intégration multimodale :

> « *Perhaps nowhere is the convergence of modalities more evident than in the superior colliculus*[18]. »

18. « La convergence des modalités est sans doute la plus manifeste dans le colliculus supérieur »

Le CS est situé au niveau du tronc cérébral et est organisé en sept couches réparties en deux unités fonctionnelles, l'une recevant des afférences sensorielles, l'autre générant des commandes motrices à partir de ces afférences. Plus qu'une intégration multimodale, le CS est considéré comme une entité d'intégration sensorimotrice, couplant ainsi la convergence d'informations sensorielles avec le déclenchement d'actions motrices ciblées vers des entités sensorielles d'intérêt. Les neurones du CS intègrent les informations spatiales afférentes provenant de la vision, de l'audition et de la proprioception. Cette intégration aboutit à la génération d'efférences motrices vers les muscles des yeux, du cou et du corps [107]. Parmi toutes les actions motrices que le CS peut générer, les plus importantes sont les saccades oculaires [108] et les mouvements de tête [109]. Par cette capacité à engendrer une réponse motrice directement fonction d'entrées sensorielles, le colliculus supérieur est aussi considéré comme une structure cérébrale impliquée dans les phénomènes attentionnels, particulièrement de type exogène (cf. **Sec. 2.3**). Plus récemment, le système vestibulaire a lui aussi été caractérisé comme étant un élément important de l'intégration multimodale et du contrôle moteur [110].

L'intégration d'informations multimodales est régie par deux principes fondamentaux : (i) si deux stimuli cross-modaux se chevauchent suffisamment spatialement et temporellement, un effet synergique sera observé dans les neurones multimodaux du CS et (ii) cette synergie sera plus prononcée lorsque la modalité des stimuli est la moins pregnante chez les neurones du CS, phénomène nommé *amélioration multimodale*. D'autre part, l'intégration multimodale est dépendante de la *congruence* des stimuli perçus, c'est-à-dire lorsque deux, ou plusieurs, stimuli sont issus de la même entité perceptive, comme un objet audiovisuel par exemple, ou partagent des caractéristiques communes (comme un clic sonore). En effet, lorsqu'il y a un conflit, par exemple, entre la vision et une autre modalité, la vision l'emporte généralement : il s'agit du phénomène qualifié de *capture visuelle* par John C. Hay *et al.* en 1965 [111]. De plus, les informations provenant de la modalité conflictuelle n'a que peu, voire pas, d'effet sur la perception visuelle : la position visuelle d'un objet n'est pas altérable par un stimulus audio spatialement incongru, comme l'a montré Herbert L. Pick *et al.* en 1969 [112]. Il existe cependant quelques cas dans lesquels l'audio prend le dessus sur la vision. En 1959, J. W. Gebhard & G. H. Mowbray [113] ont étudié l'influence de la perception de la vitesse d'un flash en fonction de la vitesse de battements sonores. Les résultats ont montré que dans ce cas, seule la modalité audio influence la perception de la vitesse du clic visuel, et non l'inverse. Ce phénomène de capture auditive a été étudié par la suite, jusqu'à ce que Robert B. Welch *et al.*, en 1980 [114] émette cette hypothèse, qui tente d'expliquer la différence entre capture audio et visuelle : la vision serait particulièrement adaptée à l'analyse spatiale tandis que l'audition serait particulièrement adaptée à l'analyse temporelle. Cette distinction peut s'expliquer par la nature même des systèmes perceptifs : le système auditif (cf. **Sec. 2.2.1**) doit traiter des informations intrinsèquement temporelles et possède des structures dédiées à la détection de différences temporelles ; la vision quant à elle, est naturellement plus précise pour la localisation spatiale grâce à la nature des stimuli visuels et la composition en cellules sensitives de la rétine. Cette hypothèse, appelée *modality appropriateness*, a alors permis d'expliquer la prédominance du phénomène de capture visuelle jusqu'alors : la plupart des expériences d'intéraction cross-modales requéraient d'effectuer une tâche de détermination d'une ou plusieurs caractéristiques spatiales (position, orientation,

forme...).

En 2001, Robert Fendrich & Paul M. Corballis [115] ont créé un paradigme expérimental permettant d'explorer plus en profondeur les phénomènes de capture audio et/ou visuelle, sur la base des travaux de Welch *et al.* Toujours utilisant des flashes lumineux et des clics sonores, les résultats obtenus montrent que lorsque deux stimuli sont temporellement proches, une capture cross-modale peut survenir, et que cette capture tend à unifier leur perception respective. Leur résultats mettent en avant un effet plus important de la capture audio que de la capture visuelle. Les causes d'une dominance particulière de l'audition ou de la vision sur l'analyse d'une scène multimodale n'étant pas suffisamment claire selon les auteurs — le phénomène de capture audio qu'ils ont observé n'étant peut-être qu'un cas de capture temporelle de phase — ceux-ci ont introduit la notion d'*Intersensory Temporal Locking*[19] (ITL) permettant de décrire plus généralement les phénomènes d'interactions multimodales. L'ITL, étayé par une étude antérieure de C. R. Scheier *et al.* en 1999 [116], est compris comme un phénomène permettant de résoudre des ambiguités temporelles dans la perception de stimuli multimodaux.

D'autres travaux de recherche ont également mis en avant l'influence de l'audio sur la vision. Notamment, Ladan Shams *et al.*, en 2001 [117] et 2002 [118], ont montré que lorsqu'un flash lumineux est accompagné de deux ou plusieurs bips audio, ce flash lumineux est alors perçu comme plusieurs flashes. Les paradigmes expérimentaux utilisés et les résultats obtenus par les auteurs permettent d'éliminer d'éventuelles causes du phénomène de capture audio : amélioration attentionnelle, mouvements oculaires, biais cognitifs ou influence cognitive descendante. Le fait que l'influence bimodale n'intervienne que dans le cas de démultiplication du flash visuel et non dans les cas de l'illusion de fusion de deux flashes en un seul, ne peut pas être expliqué par l'hypothèse de *modality appropriateness*. Les auteurs supposent donc que ces interactions cross-modales sont plutôt dépendantes des caractéristiques intrinsèques des stimuli. Cette assymétrie entre l'altération intermodale de la perception audio-visuelle a aussi été observée dix ans plus tôt, par Helena Saldaña & Lawrence D. Rosenblum [119], mais dans le cas de la vision altérant l'audio. Combinant ces deux études (et celles sur lesquelles elles reposent), et mettant en évidence le fait que la notion de continuité des stimluli était certainement la caractéristique discriminante dans les expériences conduites, Ladan Shams et al. [118] proposent l'explication de l'altération multimodale suivante :

> « *The discontinuous stimulus in one modality alters the percept of the continuous stimulus in the other modality and not as strongly vice versa*[20]. »

De nombreux modèles du CS ont été développés depuis les années 70, notamment à partir du modèle Robinson [120]. Cette structure est fondamentale et centrale dans le processus d'intégration multimodale mais aussi dans les processus attentionnels, par sa capacité à générer des actions motrices, comme nous le verrons à la **Sec. 2.3**. Cependant, par souci de concision, nous avons choisi de ne présenter qu'un seul modèle d'intégration multimodal, modèle très récent (2014) et proche des ambitions du modèle HTM. La section suivante est donc dédiée à la description de ce modèle.

19. Verrouillage temporel intersensoriel.

20. « Le stimulus discontinu d'une modalité altère la perception du stimulus continu de l'autre modalité de façon plus prononcée que le phénomène inverse. »

2.2.4.2 Application Robotique

En 2014, Kuniaki Noda *et al.* [121] ont développé un modèle d'intégration multimodale basé sur des réseaux de neurones profonds et intégré à la plateforme robotique Nao [21]. Ce modèle a été développé avec les ambitions suivantes, très proches des nôtres :

- Implémenter une mémoire cross-modale ;
- Reconnaissance robuste au bruit grâce à la capacité de généralisation des caractéristiques mutlimodales ;
- Aquisition de la causalité multimodale et capacité de prédiction sensorimotrice sur la base de cette causalité.

L'idée est d'apprendre des séquences temporelles multimodales organisées en tuples de vecteurs de caractéristiques : données odométriques (angles des jointures des bras du robot), caractéristiques extraites de l'image captées par les caméras du robot et caractéristiques extraites du son capté par les microphones. Le lien sensorimoteur permet de déclencher une réaction motrice à partir de la perception de certains objets audiovisuels cibles situés dans l'environnement. Ainsi, à cet apprentissage est lié une tâche motrice qui sera déclenchée par le contenu audiovisuel de l'environnement. Le modèle se base intensivement sur des réseaux de neurones profonds (DNN) dont l'avantage réside dans leur forte capacité de généralisation, notamment dans le cas de données manquantes. Les DNN employés ici sont entrainés pour que la couche de sortie puisse reconstruire les données envoyées à la couche d'entrée. Deux types de reconstructions sont effectuées : une basée sur le seul vecteur de caractéristiques issu de la concaténation des données sensorielles, une autre ayant pour but de prédire une séquence temporelle.

Deux expériences ont été conduites afin de valider l'algorithme de Noda *et al.* : une tâche de manipulation d'objet, tâche impliquant la vision et la manipulation, et une tâche de « *bell-ringing* », impliquant les trois modalités. Les résultats obtenus montrent globalement une bonne capacité de fusion cross-modale permettant d'inférer des données manquantes à partir de celles disponibles. Ce modèle est particulièrement intéressant pour nous puisqu'il tente de créer une représentation multimodale d'entités perçues par un robot humanoïde, dans un environnement réaliste et en ligne. De cette représentation, le robot Nao apprend une série de comportements en rapport avec la tâche à effectuer.

Cependant, plusieurs remarques sont à faire. D'une part, lors de la tâche de « *bell-ringing* », la performance de reconstruction/inférence d'image à partir du son et des données odométriques est faite sur seulement deux positions spatiales. Et sur ce scénario assez restreint, le taux moyen d'erreur d'inférence d'une image à partir du son de la cloche ainsi que des données odométriques ne descend pas en-dessous des 10%, taux relativement fort étant donné la complexité du scénario (trois objets à distinguer).

De plus, malgré l'élégance et la pertinence de l'architecture présentée, celle-ci semble lourde en regard des tâches d'intégration multimodale et d'inférence de données

21. `http://www.aldebaran-robotics.com/Downloads/Download-document/192-Datasheet-NAO-Humanoid.html`

manquantes à effectuer : trois réseaux de neurones profonds, chacun nécessitant une étape d'optimisation (de type *Hessian-free* [122]). Du point de vue de l'utilisation en ligne de cet algorithme, bien que les expériences aient été effectuées dans un environnement réaliste, avec de vrais objets et dans un temps court, l'étape d'apprentissage direct (par manipulation du bras du robot par un expérimentateur) ainsi que la nécessité d'enregistrer 10 à 20 s de données avant de pouvoir effectuer l'apprentissage rendent ce modèle incompatible avec nos contraintes. De plus, il leur a été nécessaire de fournir au réseau des données rendues unimodales artificiellement (par ajout de bruit conséquent dans les autres modalités) afin de permettre au réseau d'apprendre explicitement les corrélations entre les modalités. Enfin, les capacités prédictives de l'algorithme sont sensibles à la quantité de bruit introduite, jusqu'à rendre impossible l'inférence d'image. Les auteurs indiquent penser à ajouter un mécanisme de suppression des informations provenant des modalités dégradées, mais sans préciser comment détecter qu'une modalité est justement dégradée.

2.2.4.3 Discussion

Dans cette section nous avons présenté les mécanismes neuronaux à l'origine de la fusion de données issues des différentes sources d'informations dont l'humain dispose, notamment la vision et l'audition. Ces deux modalités sont les plus utilisées dans la formation d'une représentation interne de l'environnement riche et robuste en cela que la vision possède une grande précision couplée à une extrême rapidité dans l'analyse des données visuelles, tandis que l'audition possède, par construction, une faculté indispensable : celle de percevoir des informations issues de toutes les zones de l'environnement. La fusion multimodale est indispensable en cela qu'elle permet de faire émerger la notion d'objet, ou d'entité perceptuelle.

Nous avons également présenté une tentative innovante et récente d'implémentation robotique d'intégration multimodale d'informations auditives, visuelles et sensorimotrices. Les travaux de Noda *et al.* sur le robot Nao, proches de ce que nous cherchons à réaliser, ont permis d'entrevoir l'état de la recherche sur le problème d'intégration multimodale appliquée à un robot dans un environnement réaliste et sans connaissances *a priori* de l'environnement. Malgré une architecture puissante, nous avons mis en avant plusieurs limitations de ce modèle, motivant par la même occasion certains de nos choix concernant le modèle HTM. Une des conclusions principales que nous avons tiré des travaux de Noda *et al.* est que l'émergence de processus complexes comme l'intégration multimodale et les comportements associés à cette représentation multimodale de l'environnement est une tâche difficile à accomplir en ne prenant en compte que les caractéristiques bas-niveau des signaux perçus (données odométriques, vision et audition, pour ce modèle). Malgré l'utilisation de plusieurs réseaux de neurones profonds, d'une formalisation complexe et de larges bases de données d'apprentissage, certains résultats obtenus restent trop faibles, selon nous, pour que cette approche soit convainquante dans son ambition d'effectuer une intégration multimodale pertinente. A l'inverse, nous pensons qu'une étape d'*interprétation des données brutes*, c'est-à-dire le fait de passer sur le plan sémantique/symbolique, est indispensable à la création d'un comportement robotique basé sur une intégration multimodale des données sensorielles.

2.2.5 Conclusion

Cette section, dédiée à la *Perception*, a tout d'abord passé en revue les mécanismes de l'audition, de la vision et de leur intégration multimodale. Les systèmes auditifs et visuels présentent de nombreuses ressemblances quant à l'organisation hiérarchique capteurs/centre d'analyse. Cependant, ces ressemblances ne sont qu'apparentes : le monde visuel et le monde auditif possèdent des caractéristiques fondamentalement différentes, notamment du point de vue de leur temporalité. En effet, dans le monde audio, une intégration temporelle des informations perçues est indispensable. D'un autre côté, le monde visuel, par essence, possède une information spatiale bien plus précise et robuste. Les systèmes auditifs et visuels reflètent cette différence majeure. L'intégration multimodale de ces informations sensorielles met également en évidence les différences qui peuvent exister entre vision et audition. Bien que les mécanismes neuronaux ne soient pas encore pleinement compris et mis en évidence, nous avons présenté les nombreuses expériences ayant permis l'observation des phénomènes de capture de l'information audio ou visuelle. Cependant, ces manifestations de processus d'intégration multimodale doivent être tempérées par les travaux de recherche ayant observé des phénomènes contradictoires. Notamment, l'hypothèse d'*Intersensory Temporal Locking* de stimuli cross-modaux étant plus en faveur d'une intégration multimodale dépendante de caractéristiques intrinsèques des stimuli perçus que des modalités elles-mêmes.

D'autre part, la Théorie de la Hiérarchie Inverse a été décrite, proposant un modèle expliquant la façon dont les informations sensorielles pourraient être analysées par les aires auditives et visuelles. Cette théorie stipule que le degré d'analyse d'une information dépend de son contexte, c'est-à-dire de la facilité avec laquelle cette information est analysable : plus une information est claire pour les aires sensorielles, au sens de sa discriminabilité avec d'autres sources d'informations par exempl, moins il est nécessaire de la traiter en profondeur. Ce principe a guidé l'élaboration du modèle HTM en cela qu'il pose les bases bio-inspirées — qui seront complétées ensuite par une base plus mathématique — de la justification future de la nécessité d'inhiber les mouvements de tête préalablement générés par le modèle. En effet, le modèle HTM est un système dynamique et évolutif dans le temps : des mouvements de tête ne seront pas générés automatiquement en fonction de l'apparition d'un objet audiovisuel mais en fonction du taux d'information qu'il porte. Plus un objet audiovisuel est connu par le système, moins il est nécessaire d'acquérir de l'information supplémentaire le concernant (pouvant également être interprété à la lumière de la théorie de l'Information de Shannon [123]) rendant un mouvement de tête vers lui inutile.

Enfin, nous avons mis en avant l'importance et la prédominance de l'intégration des informations issues des différentes modalités sensorielles afin de créer une représentation complexe des entités présentes dans l'environnement. Cette intégration est majeure pour nous : elle justifie tout le travail effectué sur la conception d'objets multimodaux au cours de l'exploration d'un environnement inconnu, ainsi que l'importance de tenter de retrouver la représentation complète d'un objet lorsqu'une source d'information est manquante (objet non visible par exemple). De plus, le colliculus supérieur en tant que structure responsable de cette intégration multimodale autant que de la génération de commandes motrices vers les yeux, le cou et le corps,

est très proche d'une des ambitions du modèle HTM : générer des mouvements de tête à partir d'informations multimodales intégrées. Cette réaction motrice à des afférences sensorielles est une forme de mécanisme attentionnel. La section suivante est donc dédiée à la présentation de la notion d'*Attention*.

2.3 Processus attentionnels

Q. Qu'est-ce que l'Attention ? et comment la modéliser ?

L'Attention est, comme la perception, une notion très large regroupant un ensemble vaste de concepts et de processus. D'après William James, en 1890[124] :

> « *Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought*[22]*.* »

L'attention peut ainsi être définie comme la réquisition concomittante des « capteurs » (oreilles et yeux par exemple) vers une « entité » d'intérêt, parmi plusieurs, et des capacités d'analyse des informations conséquemment perçues. Cette entité peut être une zone spatiale, un événement sonore, l'apparition d'un objet ou encore un état interne particulier provoquant une réaction. L'attention peut être déclenchée par les caractéristiques intrinsèques des signaux : il s'agit alors d'un phénomène ascendant (« *bottom-up* ») défini alors comme « *stimulus-driven* » [125], *actif* [124] ou encore *exogène* [126]. Elle peut aussi être déclenchée par une volonté résultant d'une analyse haut-niveau, cognitive, aboutissant à la sélection consciente des signaux d'intérêt : il s'agit alors d'un phénomène descendant (« *top-down* »), qualifié de « *goal-driven* » [125], *passif* [124] ou *endogène* [126].

Les processus attentionnels ascendants sont une conséquence de la *Saillance* des caractéristiques bas-niveau des signaux perçus (cf. **Sec. 2.3.1.2**). Un exemple couramment utilisé est celui d'une image contenant un rond vert (défini comme le stimulus cible) au milieu d'un ensemble de ronds rouges (définis comme les distracteurs) : l'attention sera portée en premier lieu sur le rond vert, présentant une singularité locale (comme définie par Anne Treisman & Garry Gelade en 1980 [127]). A l'opposé, les processus attentionnels descendants motivent l'exploration spatiale ou sémantique dans un but déterminé. Par exemple, isoler puis suivre la mélodie d'une flûte traversière au sein d'un orchestre symphonique est un processus attentionnel descendant en cela que les aires corticales sensorielles vont essayer de distinguer une seule entité sonore parmi plusieurs, volontairement. L'attention, descendante comme ascendante, peut, par ailleurs, être à l'origine de mouvements : saccades oculaires et mouvements de tête notamment — la vision, en collaboration avec l'audition, étant la source d'information principale, la plus robuste et la plus rapide [94]. En conséquence de ces mouvements, de nouvelles informations vont être perçues, que ce soit la détection d'une nouvelle entité ou la mise à jour des informations concernant l'entité précédente. A leur tour, ces informations peuvent aboutir à un nouveau mouvement ou à une mise à jour de l'état cognitif et ainsi de l'état attentionnel. Il s'agit donc d'une boucle de rétroaction dans laquelle les informations requises par la partie attentionnelle vont moduler une réaction future.

22. « Tout le monde sait ce que l'attention est. Il s'agit de la prise de possession par l'esprit, de façon claire et vive, d'un objet parmi plusieurs, d'une suite de pensée parmi plusieurs. »

La partie traitant des motivations à l'exploration pourrait être intégrée à cette section portant sur l'Attention. En effet, la modélisation d'une sensibilité particulière pour un type d'entité peut être vue comme un processus attentionnel. Cependant, nous avons préféré l'intégrer à la partie traitant de l'exploration en elle-même car les paradigmes employés insistent sur le fait qu'il ne s'agit pas d'attention mais bien d'exploration. D'autre part, beaucoup de processus attentionnels sont basés sur une intégration multimodale préalable des informations sensorielles perçues. Par exemple, une incongruence entre un stimulus audio et un stimulus visuel peut provoquer une réaction attentionnelle. Ainsi, la section précédente, portant précisément sur les mécanismes de l'intégration multimodale, contient déjà de nombreux éléments qui pourraient entrer dans la description des phénomènes attentionnels : les structures cérébrales telles que le colliculus supérieur et le système vestibulaire, les notions de capture visuelle ou audio, le phénomène de *modality appropriateness* etc. Cependant, là où la section précédente ne traitait que la façon dont les informations multimodales sont rassemblées et traitées par le cerveau humain, cette section tente de décrire la façon dont l'organisme va réagir en fonction de cette intégration multimodale.

Ainsi, cette section parcourra différents volets des processus attentionnels : les bases neurales de l'attention seront décrites, avec une attention particulière sur le phénomène de « *Mismatch Negativity* » et le principe de Saillance d'une entité au sein d'un contexte perceptif. Un certain nombre d'applications robotiques basées sur le principe de saillance seront ensuite présentées.

2.3.1 Bases neurales de l'attention

2.3.1.1 Réseaux neuroanatomiques

Selon une hypothèse émise par Steven E. Petersen & Michael I. Posner datant de 1990 [128] et réévaluée en 2012 [129], le système attentionnel humain est caractérisé par trois concepts primordiaux :

1. le système attentionnel est anatomiquement séparé des systèmes analytiques (aires sensorielles et intégratives) ;
2. l'attention est organisée selon des réseaux d'aires anatomiques impliquant des structures cérébrales distinctes et des moyens de communication dédiés (en terme de neurotransmetteurs) ;
3. ces aires anatomiques ont différentes fonctions qui peuvent être caractérisées selon une approche cognitive, c'est-à-dire dont les rôles peuvent être interprétés selon des concepts haut-niveau.

Petersen & Posner mettent en avant trois réseaux neuronaux associés à de nombreuses structures anatomiques et neurotransmetteurs aux actions spécifiques [23], chacun de ces réseaux étant dédié à une forme d'attention :

alerting network (AN) permettant de réquisitionner l'attention, « être alerte », c'est-à-dire se préparer à l'arrivée d'un stimulus inattendu et pouvoir ainsi

23. Les aires cérébrales concernées ainsi que les neurotransmetteurs ou molécules associés à ces réseaux ne sont pas détaillés ici pour des raisons de concision et de pertinence.

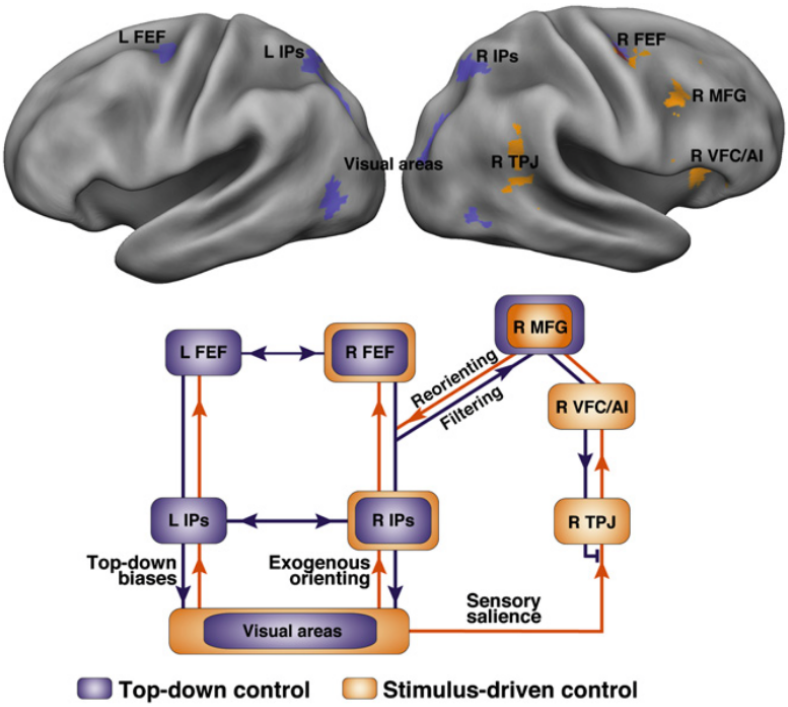


FIGURE 2.15 – RÉSEAU VENTRAL ET RÉSEAU DORSAL — (HAUT) Résultats d'une méta-analyse des régions cérébrales impliquées dans l'attention endogène (*bleu*) et exogène (*orange*). (BAS) Modèle de l'interaction entre le réseau dorsal (*bleu*) et ventral ((*orange*)) lors du phénomène de réorientation exogène. Les régions dorsales restreignent l'activation ventrale par filtrage des informations envoyées aux aires visuelles afin de promouvoir les stimuli d'importance (figure d'après [1]).

---

l'analyser plus rapidement. La voie de la norépinéphrine inclut majoritairement le cortex frontal et aires pariétales ;

orienting network (ON) permettant de prioriser une afférence sensorielle par sélection d'une modalité particulière ou d'une localisation spatiale ;

executive network (EN) lié à la détection d'une cible de l'attention. En revanche, il ne s'agit pas là d'un réseau dédié à la détection à proprement parler d'une cible d'intérêt mais plutôt du fait que lorsqu'une cible de l'attention émerge, elle capture l'attention d'une façon spécifique. En effet, dans une tâche de surveillance de plusieurs éventuelles cibles attentionnelles, lorsqu'une d'entre elles est détectée, un effet de ralentissement du temps de détection d'autres cibles est observé.

Nous souhaiterions développer particulièrement le réseau dédié à la priorisation d'une modalité ou d'une position spatiale (*orienting network*, ON). En effet, les réseaux AN et EN quant à eux n'entrent pas dans le cadre du modèle HTM : le premier est lié à l'analyse bas-niveau des stimuli — au niveau des caractéristiques des signaux audio et visuels auquel le modèle HTM n'as pas accès — ; le second impliquant l'intégration d'autres formes d'attention et de mécanismes de perception tâches-dépendantes ainsi qu'une modification du décours temporel des actions du système, actions que le modèle HTM ne peut pas non plus effectuer.

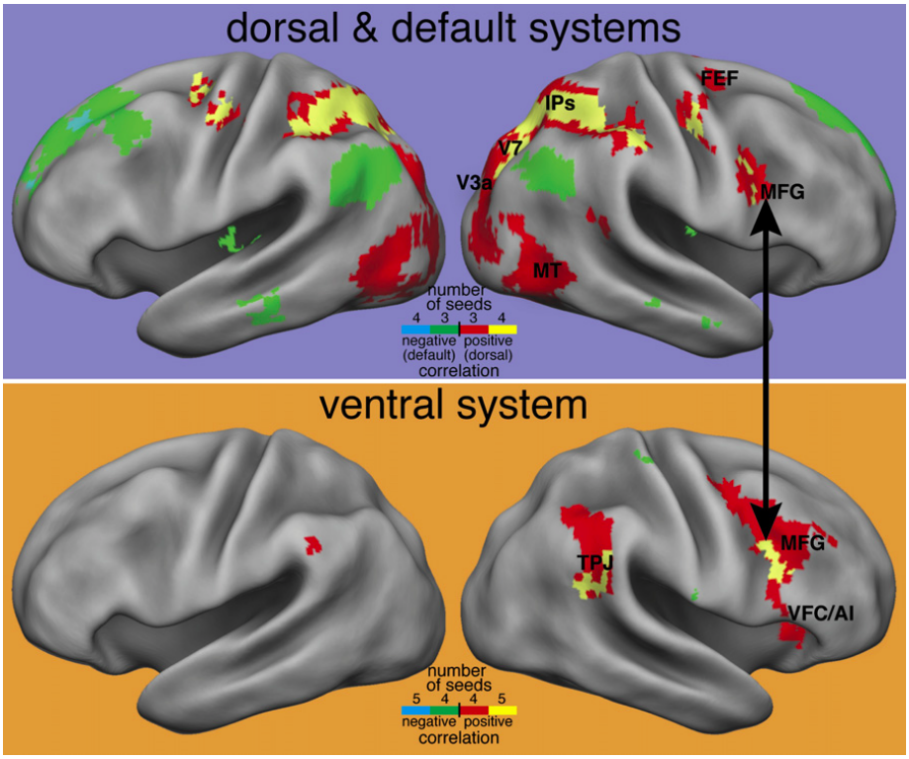


FIGURE 2.16 – RÉORIENTATION DE L'ATTENTION SELON CORBETTA *et al.* — (*haut*) L'attention portée sur un objet fait appel aux régions frontopariétales dorsales et désactive les régions ventrales ; (*bas*) Si un événement inattendu et important nécessite une réorientation de l'attention, le réseau frontopariétal ventral s'active et permet, de façon concertée avec le réseau dorsal, de modifier le comportement attentionnel (figure d'après [1]).

En 2002, MAURIZIO CORBETTA & GORDON L. SHULMAN [130] ont proposé une organisation neuroanatomique composée de deux réseaux frontopariétaux, l'un ventral (RV) et l'autre dorsal (RD) afin de modéliser le phénomène d'orientation attentionnelle. Le réseau dorsal (cortex dorsal pariétal, sillon intrapariétal, lobule pariétal supérieur et sillon précentral, cf. **Fig. 2.15**, en bleu) serait responsable du mécanisme de contrôle descendant de l'attention, i.e. de l'attention endogène. Ce réseau est capable de maintenir l'attention endogène vers un stimlus particulier sur la base du but momentané à accomplir ainsi que sur les informations précédemment aquises. Le RD envoie des signaux selon une voie descendante promouvant l'analyse des stimli d'intérêt par les aires sensorielles dédiées. Il est en fait pré-activé par différents contextes comme la prédiction de voir un objet à une position ou avec des caractéristiques particulières, par exemple dans le cas d'un mouvement oculaire [131, 132] ; ou la mémoire à court-terme d'une scène visuelle [133, 134]. En l'absence d'apparition de stimulus inattendu (cf. **Fig. 2.16**, haut), seul le RD est activé.

Le réseau ventral (jonction temporopariétale impliquant globalement les sillon et gyrus temporal supérieur, gyrus supramarginal, cortex ventral frontal, cf. **Fig. 2.15**, en orange), quant à lui, est impliqué dans l'attention exogène, c'est-à-dire celle déclenchée par les stimluli eux-mêmes. En particulier, ce réseau est activé lorsque des stimuli importants apparaissent à des positions inattendues. Son activation ne nécessite en revanche pas l'inhibition des régions frontopariétales dorsales : les deux

réseaux sont actifs conjointement afin de permettre la réorientation attentionnelle requise vers le nouvel événement. Le RV, de façon intéressante, est plus sensible à la pertinence comportementale de l'apparition d'un nouveau stimulus, plutôt qu'à sa saillance. Ainsi, selon Corbetta *et al.* [1] :

> « *The ventral network is activated by* important *stimuli that reorient attention*[24]. »

Une distinction est faite ici entre un stimulus présentant une singularité locale, du point de vue de ses caractéristiques intrinsèques (voir plus loin **Sec. 2.3.1.2**), étant donné l'environnement informatif dans lequel il est situé, et un stimulus présentant des caractéristiques sémantiques d'importance pour l'état interne de l'organisme : intérêt, récompense, danger etc. Selon cette différence faite entre saillance et importance, il est possible de caracétériser un stimulus comme saillant et important, saillant mais non important ou important mais non saillant. Ainsi, bien qu'une hypothèse en faveur de l'activation du RV par des stimuli saillants mais non importants ait été étayée par des données expérimentales en 2000 par Jonathan Downar *et al.* [135] (notamment du fait que le RV, dans des conditions passives soit activé par des stimuli distinctifs indépendemment de la modalité considérée), elle a été rapidement rejetée cinq ans après par Michelle J. Kincade *et al.* [136] et particulièrement par Iole Indovina & Emiliano Macaluso en 2007 [137]. Ces derniers ont montré que le RV (toujours conjointement avec le RD) était activé par des stimuli importants mais non saillants tandis que des stimuli très saillants mais non pertinents n'occasionnaient que peu de réponses des régions frontopariétaux ventraux. Ce comportement peut être expliqué par la nécessité de conserver son attention sur des tâches d'importance et d'être ainsi capable de filtrer des stimuli potentiellement distracteurs par leur absence de pertinence dans le contexte considéré (cf. **Fig. 2.15**, bas, « *filtering* »). Les résultats obtenus par Gordon L. Shulman *et al.* [138] en 2003, et J. Jay Todd *et al.* [139] en 2005 ont démontré que lorsqu'un sujet porte son attention sur une tâche à effectuer, les stimuli pertinents pour la tâche en question inhibent le RV, permettant ainsi d'empêcher la réorientation attentionnelle vers des objets non importants. Le réseau frontopariétal ventral semble être ainsi fortement impliqué dans la tâche de réorientation attentionnelle (i) causée par des stimuli d'importance et (ii) lorsqu'une tâche attentionnelle est déjà en cours. Ceci explique les résultats obtenus dans des conditions donc passives, lorsqu'aucune tâche attentionnelle n'est donc requise, montrant que le RV peut s'activer par l'apparition de n'importe quel stimulus saillant ou inattendu. D'autre part, cette hypothèse d'une organisation constituée de deux réseaux neuronanatomiques frontopariétaux semble être valable pour d'autre modalités que la vision : des résultats similaires à ceux présentés par Corbetta et al. *et al.* ont aussi été obtenus pour les modalités auditive et tactile [135].

L'attention exogène fait appel à la nature même des stimuli perçus par un individu. De nombreuses caractéristiques peuvent déclencher une réaction attentionnelle, étant donné un environnement, un contexte. Cependant, il est possible de généraliser les causes de déclenchement d'une réaction attentionnelle grâce à la notion de *Saillance* d'une entité en fonction de son contexte temporel et sémantique. La section suivante est dédiée à la description de cette notion.

24. « Le réseau ventral est activé par des stimuli *importants* qui réorientent l'attention. »

2.3.1.2 La Saillance

Selon le Larousse, quelque chose est dit *saillant* lorsqu'il :

1. Fait saillie, dépasse, déborde sur la ligne normale.
2. Ressort sur le reste ; attire l'attention.

La saillance est donc lexicalement définie comme une inconsistance par rapport à la normale, un trait qui s'en écarte, une fluctuation extra-ordinaire au sein d'une séquence régulière — tout au moins prédictible. Chez l'humain, la saillance intervient particulièrement dans le champ de la perception sensorielle, notamment audio et visuelle. Dans ce contexte, un stimulus est *saillant* lorsqu'il se démarque des autres et qu'il entraîne une réaction comportementale, qu'il « attire l'attention ». Par exemple, Hans-Christoph Nothdurft, en 2006 [140], pose la question suivante :

> « *Why don't we see the tree in the forest but do see the single tree in the garden*[25] *?* »

L'entité visuelle est pourtant la même mais la réaction attentionnelle sera pourtant différente. Dans le contexte de l'analyse d'un environnement sensoriel, un stimulus est *saillant* lorsqu'il se démarque des autres et qu'il entraîne une réaction comportementale, qu'il « attire l'attention ». Du côté du monde audio, Varinthira Duangudom & David V. Anderson [141] parlent du système auditif ainsi :

> « *The auditory system is well-versed in change detection. From an auditory scene analysis perspective, older sounds that are relatively constant or unchanging tend to become background, while changes in a sound or new sounds will stand out from the background and are more salient*[26]*.* »

La saillance, bien qu'intensivement utilisée dans les communautés des neurosciences, de la psychologie, de l'ingénierie et de la robotique, reste un concept général puisque très dépendant de l'environnement ou du contexte dans lequel elle existe. Dans de nombreux travaux, la saillance est considérée comme un phénomène *pré-attentif*, comme une caractéristique d'un stimulus pouvant être à l'origine d'une réaction attentionnelle [141] :

> « *Salient sounds are defined as those sounds that can be noticed without attention [...] It is pre-attentive and deals with sounds*[27] *that grab a listener's attention*[28]*.* »

De plus, la perception de la saillance d'un stimulus est également influencée par l'apprentissage et l'expérience : un musicien saura détecter une fausse note instantanément sans même se concentrer sur son écoute, stimulus considéré comme saillant,

25. « Pourquoi ne voit-on pas l'arbre dans une forêt mais voyons l'arbre seul dans un jardin ? »

26. « Le système auditif est très enclin à la détection de changements. Du point de vue de l'analyse de scène audio, les sons plus anciens et relativement constants tendent à être en fond sonore, tandis qu'un changement sonore ou une apparition d'un nouveau son va dénoter et être ainsi considéré comme saillant. »

27. Peut également s'appliquer aux entités visuelles.

28. « Les sons saillants sont définis comme ceux qui sont détectés sans attention. C'est un mécanisme pré-attentionnel et concerne les sons qui attireront justement l'attention de l'auditeur. »

alors qu'elle pourrait passer inaperçue par une personne non formée. Il n'est ainsi pas aisé d'apporter une formalisation mathématique de la saillance mais il est possible de la définir selon deux approches : du point de vue des caractéristiques des signaux ou du point de vue sémantique. La première approche tente de comprendre quelles caractéristiques d'un signal le rendent saillant : intensité, contraste fréquentiel, couleur etc. La seconde tente d'inclure la saillance dans un contexte informationnel plus global (type *entropie de Shannon* ou théorie bayesienne).

Dans le système visuel humain, les mécanismes attentionnels basés sur la saillance aboutissent au déclenchement de mouvements moteurs, par exemple des saccades oculaires vers les stimuli de forte intensité [142, 140]. Ces saccades oculaires, manifestations visibles de la réaction attentionnelle à un stimulus visuel (« *overt attention* », par opposition aux réactions attentionnelles non visibles — cérébrales — qualifiées de « *covert attention* »), sont présentes très tôt dans le développement cérébral humain : dès la naissance, les mouvements saccadiques sont déjà relativement matures. Ces mouvements saccadiques sont contrôlés par le colliculus supérieur (CS), structure également responsable de l'intégration multimodale des stimuli sensoriels (cf. **Sec. 2.2.4**). Le CS contrôle l'amplitude et la direction de ces saccades, mais est également impliqué dans la fusion d'embryons de cartes topographiques de saillance visuelle et auditive.

Un peu plus tôt, en 1999, Mark A. McDaniel *et al.* [143] ont effectué une étude sur l'implication du lobe frontal et du lobe temporal médian dans des tâches impliquant la mémoire prospective, mémoire permettant de planifier une action dans le futur. Leurs résultats ont montré un rôle bien plus important du lobe frontal dans la tâche à effectuer. Bien que ne plaçant pas la saillance des événements directement au centre de leur paradigme expérimental, il est intéressant de noter que les auteurs ont observé une influence notable de la saillance des événements cibles dans la performance des participants.

Comme dans la plupart des processus perceptifs, la construction d'une carte rassemblant et organisant l'ensemble d'un type d'information est possible. Dans le système visuel, le cortex visuel primaire (V1) contient déjà, selon Zhaoping Li [144], une carte de saillance. En 2003, James A. Mazer & Jack L. Gaillant [145] ont montré que l'activité des neurones de l'aire visuelle extrastriée V4, structure plus haute dans la hiérarchie de l'analyse des signaux visuels, peut prédire, durant une tâche d'exploration visuelle, si une saccade oculaire va être déclenchée vers une zone spatiale particulière, observation en faveur de la présence d'une carte topographique de saillance dans V4. De plus, l'aire intrapariétale latérale [146] et le « *Frontal Eye Field* » [147] ont également été associées au phénomène de saillance visuelle.

Concernant l'environnement auditif, la saillance aboutit de façon similaire à des réactions motrices de type mouvements de tête et de corps. Cependant, les caractéristiques auxquelles le système auditif est sensible, et sur lesquelles il se base pour interpréter une scène visuelle sous le prisme de la saillance des événements s'y déroulant, sont différentes de ceux utilisés par la vision. Notamment, le système auditif analyse principalement les informations sonores par leurs modulations spectrales et temporelles [148, 76] et, sur cette base, est capable d'extraire des entités auditives d'intérêt, même au sein d'environnements bruités [149]. Les différentes caractéristiques acoustiques, notamment le contraste fréquentiel, le contraste temporel et l'in-

tensité, sont intégrées en parallèle par les neurones des aires auditives, aboutissant à la formation de cartes de saillance dédiées à la caractéristique considérée. Ces cartes sont ensuite fusionnées afin de former unex carte globale de la saillance auditive de l'environnement acoustique considéré.

Plus récemment, en 2007, Selim Onat *et al.* [150] ont montré, chez des sujets humains et au cours d'expériences multimodales, que l'attention est principalement attirée par les stimuli audio. Ainsi, les aires sensorielles sont sensibles à une forme de continuité des informations perçues, qu'elles soient temporelles ou sémantiques. La question suivante se pose alors : comment les aires sensorielles réagissent-elles, du point de vue neuronal, à une telle discontinuité ?

2.3.1.3 Implication de la multimodalité

En addition aux structures neurales responsables ou impliquées dans l'attention clairement identifiées, existent des processus mis en avant en psychologie expérimentale permettant de rendre compte de l'impact de la multimodalité et du contexte dans les processus attentionnels. Ce qui suit fait fortement écho à la section présentant les mécanismes neuronaux de l'intégration multimodale (**Sec. 2.2.4**). Des concepts déjà détaillés seront retrouvés ici mais, cette fois-ci, sous le prisme des mécanismes attentionnels, c'est-à-dire des réponses motrices consécutives à cette intégration multimodale.

Le colliculus supérieur (CS) est une structure cérébrale fortement impliquée dans l'intégration multimodale également capable de générer des commandes motrices en fonction des informations multimodales perçues. Ces commandes motrices peuvent être autant des mouvements des yeux, du cou ou du corps entier. D'autre part, les neurones du CS répondent plus fortement lorsque des stimuli multimodaux sont spatialement congrus, ce qui correspond à une forme de calcul de saillance multimodale. Terrence R. Stanford *et al.*, en 2005 [151] a tenté de quantifier cette intégration, aboutissant à la conclusion suivante : l'intégration multimodale effectuée par le CS est équivalent à une somme linéaire des réponses unimodales.

Ian P. Howard & William B. Templeton, en 1966 [152] ont mis en évidence la tendance qu'ont les informations visuelles à dominer la perception globale et la mémorisation d'un environnement sensoriel, sur la base d'hypothèses datant des années 1930 [153]. La vision peut même occulter la perception de stimuli auditifs, comme les expériences de Francis B. Colavita de 1974 [154] l'ont montré. Selon Michael Posner *et al.*, dans leur revue de 1976 sur la *dominance visuelle* [155], la raison pour laquelle la vision prend le dessus sur d'autres modalités pourrait provenir de la « relative faible capacité des stimuli visuels à alerter l'organisme de leur apparition [29] [155] ». Ainsi, l'attention est préférentiellement portée sur l'analyse visuelle afin de contrebalancer ce relatif manque de *saillance* inhérent aux stimulis visuels.

Cependant, en 2002, Massimo Turatto *et al.* [156] ont tempéré l'importance de la dominance visuelle dans les processus attentionnels multimodaux en mettant en évidence l'impact de nombreux biais méthodologiques expérimentaux, originellement

29. « *Relatively weak capacity of visual inputs to alert the organism to their occurrence.* »

listés par Charles Spence & Jon Driver dans les années 90 [157, 158, 159, 160]. Les expériences de Turatto *et al.* ont montré que l'analyse de scènes audiovisuelles et les réactions attentionnelles consécutives sont des phénomènes influencés (i) par la modalité considérée et (ii) par la localisation spatiale des stimuli. Notamment, en prenant en compte la distinction neurophysiologique communément faite entre l'espace *personnel* ou *péricutané*, l'espace *péripersonnel* et l'espace *extrapersonnel* [161], Turatto émet l'hypothèse que la localisation spatiale relative au sujet des sources audio ou visuelles a un impact sur l'allocation des ressources attentionnelles : avoir conduit des expériences multimodales en utilisant des sons transmis par des écouteurs (espace péricutané) mais des stimuli visuels transmis par un écran (espace péripersonnel) a pu avoir un impact sur l'importance accordée à la dominance visuelle.

L'attention existe grâce à un grand nombre de structures cérébrales et de processus neuronaux liés à la perception de signaux ainsi qu'à la volonté de se concentrer sur une entité distincte de l'environnement. Bien que, comme le disait William Jones dans la citation introduisant cette section, « tout le monde sait ce que l'attention est », il est difficile de la définir précisément. La variété des informations pouvant déclencher une réaction attentionnelle ainsi que la multiplicité des motivations à porter son attention sur une entité particulière font que l'*attention* est un phénomène large et éminement complexe. Cependant, nous avons choisi de mettre en avant deux phénomènes liés directement aux processus attentionnels ascendants : la *Saillance* d'un stimulus audio ou visuel et la « *Mismatch Nagativity* ». Le premier est une caractéristique du signal pouvant entraîner une réaction attentionnelle de type ascendante (du stimulus vers les aires décisionnelles), particulièrement mis en avant dans le système visuel. Le second est une réaction quasi réflexe à des stimulis imprédictibles au sein d'une séquence prédictible. Il aurait été possible de traiter ces deux phénomènes dans la description des phénomènes perceptifs, en cela qu'ils y sont également fortement liés. Cependant, la saillance et la MMN nous intéressent ici particulièrement par leur impact sur l'état attentionnel. Ces deux phénomènes sont détaillés aux paragraphes suivants.

2.3.1.4 Capacités de prédiction du cerveau

Les aires sensorielles ne sont pas des structures cérébrales uniquement destinées à analyser les informations qui leur sont envoyées puis à les transformer et les intégrer en une information plus haut-niveau. Les aires corticales auditives et visuelles sont également dotées de capacités de *prédiction* (les capteurs eux-mêmes, comme la rétine [162], font preuve de capacités de prédiction). Karl Friston, en 2005 [163], met en avant le fait que le cerveau possède des représentations internes du monde à partir desquelles il est capable de prédire ce qui peut arriver dans l'environnement sensoriel. Tim Lochmann & Sophie Deneve, en 2011 [164], parlent également de *codage prédictif* permettant d'inférer la cause d'événements inaccessibles pour les capteurs sensoriels. La prédiction d'information permet d'accélérer les processus d'analyses des signaux (comme la Théorie de la Hiérarchie Inverse le stipule) mais aussi de détecter des discontinuités sémantiques dans les signaux perçus. Par *discontinuité sémantique*, nous comprenons *événement imprédictible*. Et à cet événement imprédictible, le cerveau va générer une réaction, souvent motrice,

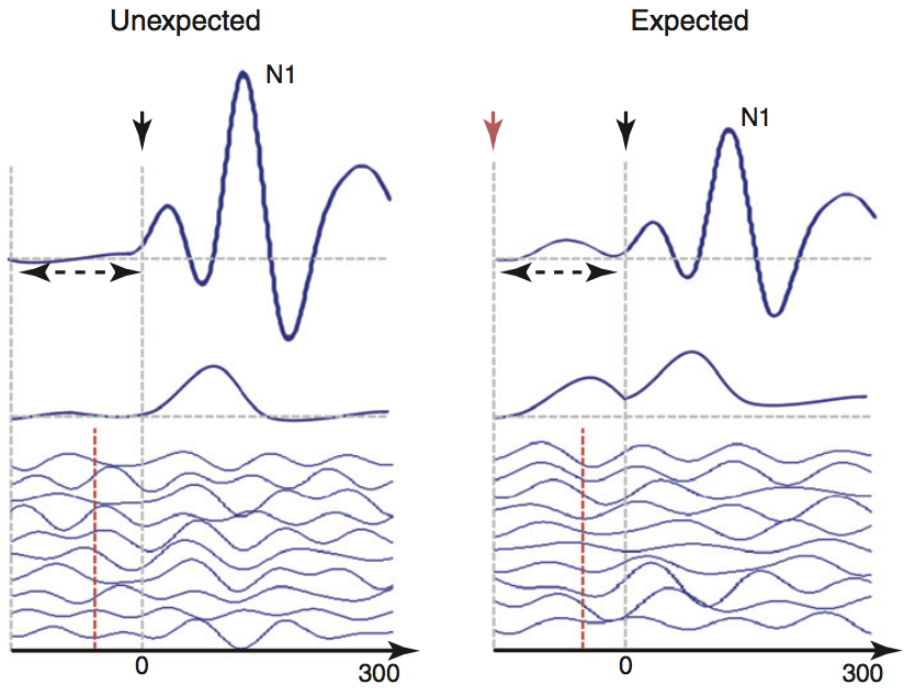


FIGURE 2.17 – CAPACITÉ DE PRÉDICTION PAR LES CORTEX SENSORIELS — Comparaison de la réponse neuronale au niveau du cortex temporal à des stimuli sonores. (*gauche*) Réponse neuronale à un stimulus imprédictible : aucune activité avant l'apparition du stimulus et forte amplitude de la réponse neuronale ; (*droite*) réponse neuronale à un stimulus prédictible : activité anticipée des neurones et diminution de l'amplitude de la réponse (faisant écho à la Théorie de la Hiérarchie Inverse, cf. **Sec. 2.2.3**) (figure d'après [165]).

permettant de confronter le monde perçu au temps t avec la prédiction émise par le cerveau au temps $t-1$. La réaction motrice est multiple : les mouvements réflexes de retrait (extension ou flexion) ont pour but d'éloigner des parties du corps pouvant être en contact avec un élément imprédictible et potentiellement dangereux ; ceux, attentionnels, ont pour but de réquisitionner les capteurs afin d'aquérir plus d'information sur l'événement imprédictible, qu'ils soient causés par la volonté du sujet — attention endogène — ou par l'événement en lui-même — attention exogène. Seul le deuxième type de réaction motrice sera considérée ici : le robot n'est pas équipé pour avoir des réactions motrices réflexes impliquant des membres du corps.

LUC ARNAL & ANNE-LISE GIRAUD, dans leur revue de 2012 sur les oscillations corticales et les prédictions sensorielles [165], regroupent les mécanismes observés permettant au cortex auditif de prédire *quand* un stimulus va arriver. Cette prédiction, évidemment, n'est pas indépendante du contexte dans lequel le stimulus prédit survient : il s'agit par exemple de prédire l'arrivée d'un stimulus au sein d'une séquence répétée ou prédictible, ou de l'apparition d'un stimulus audio en fonction d'un événement. Par exemple, lorsqu'un verre tombe par terre, il est possible d'anticiper le son que cet événement va produire.

Un peu plus tôt, nous avons détaillé les mécanismes cérébraux de l'attention, notamment par la présentation de l'hypothèse d'une organisation en deux réseaux neuroanatomiques (réseau frontopariétaux ventral et dorsal) des mécanismes atten-

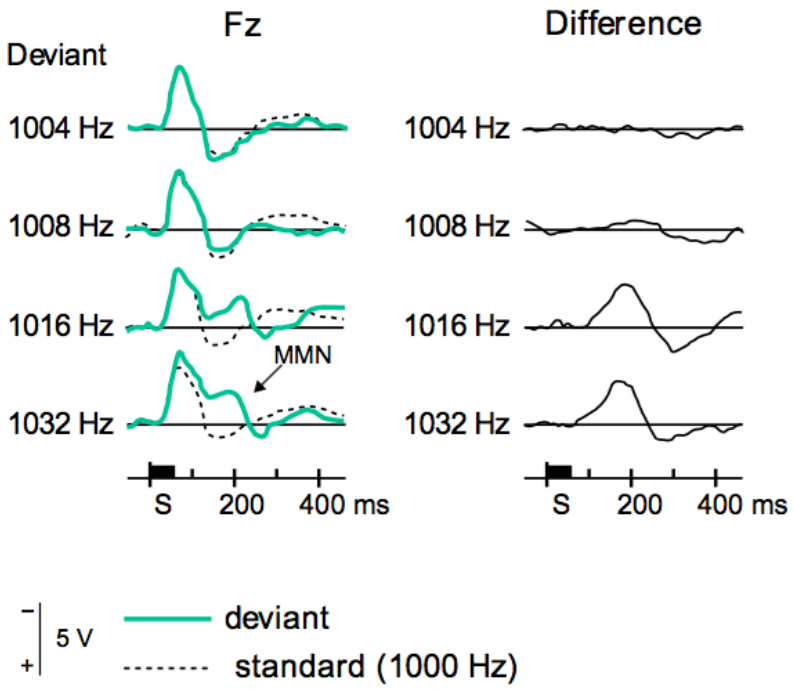


FIGURE 2.18 – MISMATCH NEGATIVITY — (*gauche*) Réponses neuronales enregistrées pour des sons à 1000 Hz présentés aléatoirement (80% des occurrences, pointillés noirs) et pour des sons déviants à différentes fréquences notées à gauche (20% des occurences, ligne verte) ; (*droite*) Différence entre réponses aux stimuli de référence (1000 Hz) et les stimuli déviants (figure d'après [166, 167]).

---

tionnels. La description de cette architecture cérébrale a souvent fait appel à la notion de stimulus *inattendu*. Cette caractéristique des stimuli perçus est justement détectable par les aires sensorielles. Et être capable de détecter quelque chose d'inattendu implique la capacité de prédire ce qui aurait dû arriver et de pouvoir le confronter avec ce qui est réellement arrivé. Plus particulièrement, une des caractéristiques des stimuli audio et visuels les plus étudiées dans le domaine de l'attention, de l'intégration multimodale et des capacités de prédiction des aires sensorielles, est la *Saillance* d'un stimulus (voir plus haut). Nous souhaitons ici introduire une manifestation neuronale présente dans les aires sensorielles et qui est une conséquence de cette réaction à l'apparition inattendue d'un stimulus ou d'une dynamique imprédictible d'un même stimulus : la « *Mismatch Negativity* » (MMN), phénomène observé et formalisé par RISTO NÄÄTÄNEN *et al.* en 1978 [168]. La MMN est une réaction à l'*incongruence* d'un événement étant donné son contexte à court-terme. Elle apparaît dans toutes les aires sensorielles du cerveau mais est particulièrement présente dans le système auditif [169] au sein du cortex temporal supérieur et du cortex frontal [170]. Par exemple, si au sein d'une une séquence répétée d'un son pur de fréquence 1000 Hz, un son de fréquence 2000 Hz apparaît, ce son sera considéré comme déviant en regard de la séquence prédictible. Egalement provoquée, dans les aires auditives, par de faibles variations d'amplitude et de timbre d'un son [171], la MMN est une réaction qui n'est pas simplement causée par l'apparition d'un nouvel objet perceptif mais également par une variation de la dynamique d'un même stimulus. Elle est observée entre 100 ms et 200 ms après l'apparition du stimulus ou plus globalement de l'événement *déviant*. La **Fig. 2.18** illustre la MMN dans le cas

de la présentation d'une séquence répétée de sons à 1000 Hz au sein de laquelle ont été introduits, de façon aléatoire, des stimuli déviants en terme de fréquence fondamentale. Cette réaction apparait aussi bien pour des stimuli simples que pour des séquences de sons de paroles, nécessitant un traitement important du signal audio. Selon R. J. Zatorre & M. Schönwiesner [172], la MMN est considérée comme un mécanisme d'alerte déclenchant une réaction comportementale attentionnelle. Apparaissant ainsi très tôt dans l'analyse des informations sensorielles et ayant des conséquences allant jusqu'à une réaction motrice, la MMN montre que la notion de Congruence d'un événement perceptif est majeure dans la compréhension d'un environnement, du point de vue des entités audiovisuelles y étant présentes.

2.3.1.5 Discussion

Nous avons ici passé en revue certains des principaux mécanismes cérébraux responsables des phénomènes attentionnels observés chez l'humain. Nous avons particulièrement détaillé les principes de l'attention exogène, causée par les stimuli perçus, opposée à l'attention endogène, causée par l'individu (tâche à effectuer, motivation à porter son attention sur une partie particulière de l'environnement ou à un type de stimuli en particulier). A noter que l'attention endogène est également liée aux motivations à l'exploration, comme présentées à la **Sec. 2.1.3** : en effet, la motivation, et particulièrement celle intrinsèque, est une volonté interne aboutissant à la sélection consciente de zones de l'environnement sur lesquelles porter son attention, indépendemment des stimuli perçus. L'attention exogène est une réaction quasi immédiate à ce qui est perçu, au point même que les aires sensorielles ne sont parfois pas requises pour générer une réaction motrice en réponse à un stimulus — comme dans le cas de la vision où environ un cinquième des informations visuelles sont envoyées directement au colliculus supérieur, structure impliquée dans une forme prématurée d'intégration multimodale et de génération de commandes motrices (cf. **Sec. 2.2.4** & **Sec. 2.3.1.3**). De plus, l'existence de réactions neuronales comme la MMN apparaissant très tôt après l'apparition d'un stimulus (entre 100 ms et 200 ms) renforce l'importance d'être capable de détecter rapidement, et avant d'avoir accès à une représentation complexe de l'entité perceptuelle, la présence d'un événement incongru, que ce soit du point de vue de son contenu sémantique que de sa dynamique temporelle attendue. Les capacités prédictives des aires sensorielles montrent également qu'elles ne sont pas que des structures analytiques passives, ne réagissant que lorsque des informations arrivent, mais qu'elles sont capables d'émettre des hypothèses sur la probabilité d'évolution d'un événement perçu. Ces hypothèses, si elles se trouvent vérifiées, permettent de diminuer le temps d'analyse des stimuli ainsi que la profondeur de ces analyses.

Nous présentons dans la section suivante certaines applications robotiques et computationnelles de phénomènes attentionnels basés sur la saillance d'un stimulus.

2.3.2 Applications robotiques

Dans le domaine robotique, la notion de *filtrage attentionnel* est souvent utilisée afin de modéliser le phénomène de l'attention, caractéristique importante des

systèmes intelligents. L'attention peut en effet être considérée comme un filtre appliqué aux informations perçues en cela qu'elle va masquer ou au contraire mettre en avant une partie restreinte de celles-ci afin de faciliter l'analyse des informations pertinentes pour la tâche à effectuer. Dans de nombreux travaux de recherche en robotique attentionnelle, la saillance est utilisée en tant que modélisation bas-niveau de l'attention. Particulièrement, la création de cartes de saillance, initiée par Cristof Koch & Shilmon Ullman en 1985 [173], est une des finalités des modèles computationnels de l'attention : ces cartes permettent de lier la saillance d'un événement avec la position dans l'environnement du système artificiel, lui permettant ainsi de décider sur quelle zone de l'environnement porter son attention, i.e. réquisitionner ses capteurs pour acquérir plus d'informations sur cet événement. Laurent Itti & Christof Koch, en 2001 [174], ont défini cinq principes majeurs autour desquels le domaine de la modélisation de l'attention visuelle s'articule :

- Saillance perceptuelle ;
- Carte de saillance permettant un contrôle ascendant ;
- Inhibition du retour permettant de limiter les minimums locaux, ce qui a pour effet de « coincer » l'attention sur une même zone de l'environnement ;
- Mouvements des yeux, saccades oculaires notamment ;
- Compréhension de la scène et reconnaissance d'objets, partie de plus haut niveau des algorithmes de modélisation de l'attention.

Cette section détaillera quelques modèles computationnels de l'attention en robotique basés sur cette notion de saillance et ayant pour but la construction de cartes cognitives permettant l'émergence de comportements attentionnels. L'inhibition du retour et la propension du cerveau à considérer l'environnement par les entités multimodales qui le composent ont été détaillées plus haut. Ces concepts sont des bases de l'attention et ne sont pas implémentés sous forme de modèles computationnels en eux-mêmes. Les quelques modèles présentés ici concernent surtout la modélisation de l'attention visuelle pour plusieurs raisons. Premièrement, le système visuel a été beaucoup plus étudié que le système auditif, notamment parce qu'intrinsèquement, une scène visuelle est plus facilement analysable qu'une scène audio, que ce soit du point de vue de la discriminabilité des entités visuelles entre elles et que du caractère beaucoup moins temporel des informations visuelles. Ainsi, la mise en place de dispositifs expérimentaux impliquant des scènes visuelles est plus simple que pour des scènes auditives. Deuxièmement, les manifestations de l'attention visuelle sont facilement mesurables, notamment *via* l'enregistrement des saccades oculaires, phénomène le plus représenté dans l'attention visuelle. Les saccades oculaires sont des mouvements extrêmement rapides et générés très souvent. Elles font partie de l'attention visible, conjointement avec les mouvements de tête. Concernant l'attention auditive, la seule manifestation visible du phénomène d'orientation attentionnelle est les mouvements de tête, mouvements plus lents, plus rares et multimodaux (sous l'effet des afférences informations visuelles, tactiles et olfactives, entre autres). D'autre part, les saccades oculaires sont des mouvements issus de l'attention exogène, *stimulus-driven*, contrairement aux mouvements oculaires plus lents pouvant être d'origine endogène. Ainsi, il est plus facile de distinguer une réaction attentionnelle causée par les stimuli perçus uniquement, mettant de côté l'éventuel impact d'une réaction attentionnelle endogène.

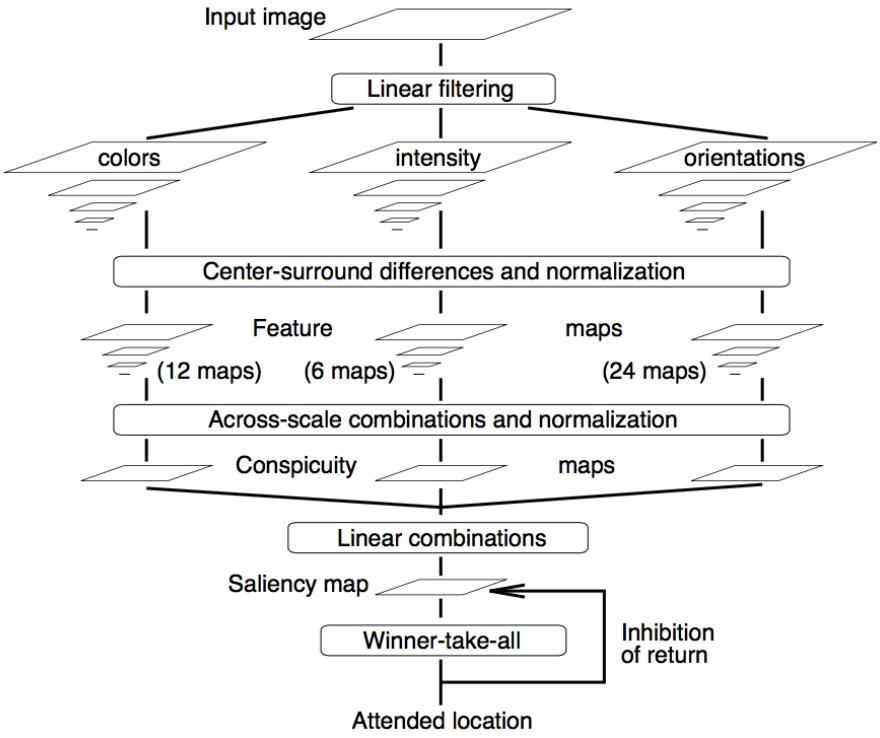


FIGURE 2.19 – CARTE DE SAILLANCE VISUELLE SELON ITTI *et al.* — A partir de l'analyse de plusieurs caractéristiques bas-niveau des signaux visuels perçus, des cartes de saillance se forment permettant, par leur combinaison, d'effectuer une analyse du point de l'espace le plus attractif du point de vue attentionnel. Ce point sera alors la cible d'un mouvement des yeux (figure d'après [175]).

2.3.2.1 Saillance visuelle

En 1998, LAURENT ITTI *et al.* [175], sur la base des travaux fondateurs de CHRISTOPH KOCH *et al.* en 1985 [173], ont proposé un modèle de carte de saillance visuelle inspirée du comportement et de l'architecture neuronale du cortex visuel primaire (V1). La première étape de la création d'une carte de saillance visuelle est l'extraction de caractéristiques : couleur, intensité et orientation. Cette extraction est représentée sous forme de carte reliant la position des entités visuelles détectées avec la caractéristique considérée. La deuxième étape correspond au traitement de ces cartes selon une méthode dite de « *center-surround differences and normalization* » imitant le fonctionnement de type « champ récepteurs » des cellules de la rétine, du noyau géniculé latéral et de V1 (mais également observé dans les aires auditives). Le principe de champ récepteur, dans les aires visuelles par exemple, consiste en une réaction activatrice lorsque le stimulus est situé dans une région précise et étroite de l'espace (*center*) et une activité inhibitrice lorsque le stimulus est situé dans une région plus large, région concentrique au centre (*surround*). La troisième étape est une combinaison linéaire des cartes résultantes, comme illustré à la **Fig. 2.19**.

Le but de cette carte de saillance visuelle est d'obtenir une quantification de la saillance des entités détectées en fonction de la position spatiale dans le champ visuel. Cette carte, et la quantification sous-jacente, permet ainsi de guider l'exploration

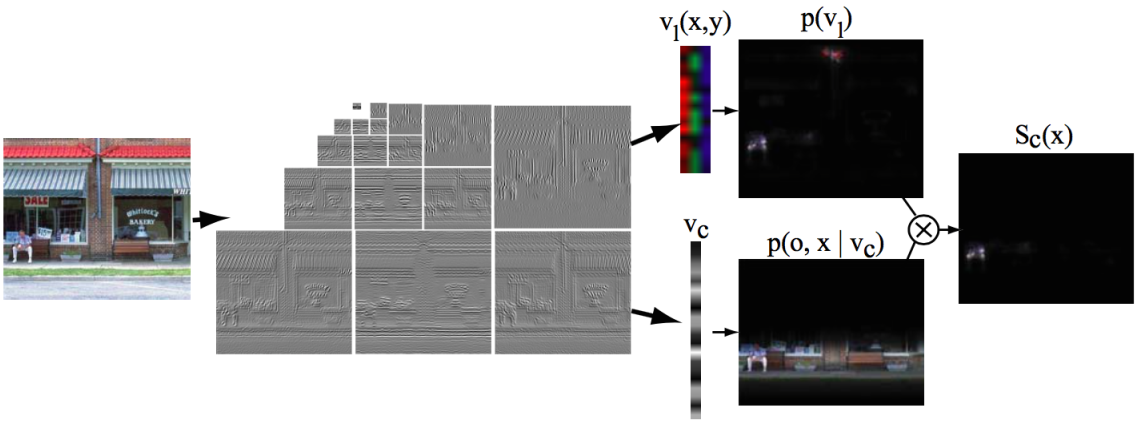


FIGURE 2.20 – CARTE DE SAILLANCE VISUELLE SELON OLIVA *et al.* — Système attentionnel basé sur une carte de saillance probabiliste (figure d'après [177]).

visuelle de la scène en considérant les points les plus saillants comme ceux nécessitant une attention particulière. Leur modèle, comparé notamment à une analyse de type *Spatial Frequency Content*[30] proposé par PAMELA REINAGEL & ANTHONY M. ZADOR [176] un an plus tôt, montre de bien meilleures performances et une robustesse bien plus grande à l'addition de différents types de bruit. De plus, ce modèle d'attention visuelle basé sur la saillance a des performances très comparables à celles mesurées chez l'homme dans des tâches d'exploration visuelle d'image fixe comme décrites par ANNE M. TREISMAN & GARRY GELADE [127]. Cependant, une des limitations de leur modèle est le nombre et la qualité des caractéristiques extraites des signaux visuels. Notamment, les auteurs insistent sur l'absence d'analyse de mouvement, supposée jouer un rôle majeur dans l'analyse de la saillance visuelle et dans l'attention, ainsi que de mécanismes récursifs permettant, entre autres, de reproduire les phénomènes de complétion/fermeture de contour, phénomènes également fortement impliqués dans certains processus d'analyse de caractéristiques visuelles.

En 2003, AUDE OLIVA *et al.* [177] ont proposé un modèle de saillance visuelle prenant en compte le contexte visuel et selon l'approche probabiliste suivante : la saillance est définie comme une faible probabilité d'observer un groupe de caractéristiques dans une image. Sur la base d'un découpage de l'image selon le principe de *pyramide orientable*[31] [178], illustré à la **Fig. 2.20**, permettant de décomposer une image en sous-bandes d'orientation et d'échelle différentes, les auteurs aboutissent à une représentation complexe d'une image. A partir de cette analyse, la saillance d'un point de l'image est définie par la probabilité de trouver l'ensemble de caractéristiques définissant ce point. L'idée est de modéliser des mouvements oculaires balayant la scène visuelle de façon similaire à l'humain. Les résutats obtenus par cet algorithme ont été comparés aux résultats obtenus par CHRISTOPH ITTI *et al.* [175], à une exploration visuelle aléatoire ainsi qu'aux mouvements oculaires mesurés chez des sujets humains. La première version de cet algorithme donne cependant de moins bons résultats que ceux obtenus par CHRISTOPH ITTI *et al.*

Mais ce qui rend les travaux d'AUDE OLIVA *et al.* particulièrement intéressants est

30. Contenu fréquentiel spatial.
31. *Steerable pyramid.*

l'amélioration apportée à leur algorithme, à savoir l'inclusion d'informations contextuelles. Selon une approche bayésienne reliant la probabilité de présence d'un objet visuel à une position donnée à son vecteur de caractéristiques contextuel défini *a priori*, la version améliorée de l'algorithme aboutit à des performances significativement meilleures que celles d'Itti *et al.* et proches de celles mesurées chez les sujets humains. Cette approche est intéressante en cela qu'elle montre l'imporance d'inclure des données contextuelles dans l'analyse de scène visuelle (et par extension, de scène audio). Une des limites, en revanche, est qu'une étape d'apprentissage supervisé est nécessaire avant de pouvoir utiliser la version bayésienne de leur algorithme.

2.3.2.2 Saillance auditive

La modélisation de la saillance audio est similaire à celle de la saillance visuelle : (i) extraction de caractéristiques et représentation sous forme de cartes, (ii) traitement des cartes (normalisation, applications de seuils etc.) et (iii) combinaison linéaire en une seule carte globale de saillance. Cependant, les caractéristiques utilisées pour l'analyse préalable des signaux perçus sont différents. Cette section décrit ainsi quelques modèles de saillance audio développés.

En 2005, Christoph Kayser *et al.* [179] ont proposé un modèle qui est devenu une des bases sur laquelle beaucoup de travaux ultérieurs sur la saillance auditive se fondent (illustré à la **Fig. 2.21**). Leur approche, introduite à la **Sec. 2.3.1.2** et inspirée des cartes de saillance visuelles proposées par [173], se base sur l'extraction de caractéristiques des signaux audio telles que l'intensité, le contraste fréquentiel et le contraste temporel. La carte de saillance globale obtenue permet de mettre en évidence des zones spectrotemporelles saillantes. En comparaison des performances humaines mesurées lors cette étude, le modèle proposé par les auteurs permet d'expliquer une partie des événements saillants détectés par les humains. Parmi les trois caractéristiques acoustiques extraites pour former la carte de saillance (intensité, contrastes fréquentiel et temporel), il est intéressant de noter que l'intensité est celle la moins discriminante chez les sujets testés. L'algorithme de Kayser *et al.* permet de modéliser la saillance audio de façon performante, en comparaison des résultats obtenus chez des sujets humains. De plus, elle est particulièrement intéressante en cela qu'elle montre que l'analyse d'une scène audio selon le principe de cartes de saillance est similaire à celle d'une scène visuelle, seules les caractéristiques extraites des signaux étant différentes. En revanche, cette approche est difficilement exploitable en temps réel puisqu'elle nécessite une certaine intégration temporelle.

En 2007, Varinthira Duangudom & David V. Anderson [141] ont proposé un modèle de carte de saillance audio, inspirée de la carte de Kayser *et al.* Les auteurs se basent notamment sur différentes caractéristiques acoustiques à extraire des sons perçus : (i) énergie globale, (ii) modulation temporelle, (iii) modulation spectrale et (iv) modulations temporelle et spectrale élevées. En addition à ce changement de caractéristiques, l'architecture proposée par les auteurs inclut une étape d'inhibition des cartes créées consécutivement à l'étape d'extraction de caractéristiques. Les performances du modèle ont ici aussi été comparées à celles d'individus humains testés sur une tâche de détection de saillance. Leurs résultats sont légèrement supérieurs à ceux obtenus par [179]. Une des explications des auteurs pour expliquer ces résultats est qu'ils n'incluent pas de processus *top-down*. Une autre

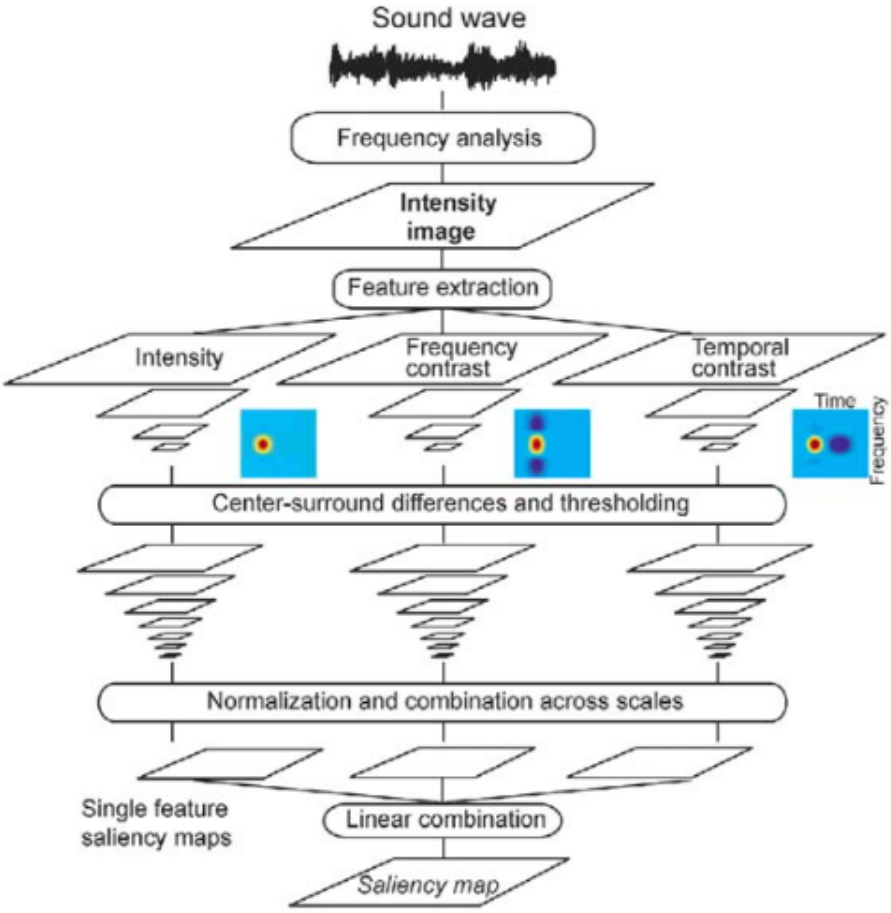


FIGURE 2.21 – CARTE DE SAILLANCE AUDITIVE SELON KAYSER *et al.* — Schéma du modèle de carte de saillance auditive. Une onde sonore perçue est convertie en une représentation spectro-temporelle permettant l'extraction de caractéristiques telles que l'intensité, le contraste fréquentiel et le contraste temporel grâce à différents filtres. Le résultat de ces extractions sont des cartes qui, après normalisation, peuvent être combinées en une carte de saillance auditive (figure d'après [179]).

---

explication possible est que la saillance n'a été définie que par les caractéristiques premières des signaux audio, sans aucune mise en contexte de ces signaux.

En 2007 également, OZLEM KALINLI & SHRIKANTH NARAYANAN [180] ont développé un modèle de saillance audio sur la même base de caractéristiques que DUANGODOM & ANDERSON et directement inspirée de ITTI *et al.* mais, sur la base d'une représentation spectrotemporale du son perçu et selon une approche en champs récepteurs (ou filtres réceptifs), ils y ajoutent une analyse de l'orientation et de la distribution de la hauteur tonale. Ainsi, cinq caractéristiques sont ici utilisées au lieu de trois. Leur modèle a été utilisé pour la détection de syllabes proéminentes au sein de sons de parole et les résultats obtenus montrent une bonne perfomance en comparaison à une base de données de sons de paroles majoritairement manuellement étiquettés pour permettre la comparaison entre l'algorithme et les performances humaines. Un des avantages de leur algorithme est qu'il est indépendant du langage et qu'il permet l'émergence d'une réaction attentionnelle selon une approche non supervisée. De plus, leur approche permet une analyse rapide des sons perçus. Ce-

pendant, les auteurs indiquent, dans la présentation de leurs travaux futurs, que les poids utilisés pour la création de la carte de saillance seront appris d'une façon supervisée afin d'être plus adaptés à la tâche à effectuer (scène acoustique générale, analyse de sons de parole etc.).

A la suite de ces travaux, Ozlem Kalinli *et al.*, en 2009 [181], ont développé un modèle de classification de données sonores après analyse de leur saillance audio : le modèle *Latent Indexing using SAliency* (LISA). Sur la base de leurs travaux précédents [180] présentés ci-dessus, ainsi que sur le concept de *Latent Perceptual Indexing* [182], les auteurs implémentent ici un algorithme combinant classification traditionnelle de sons, sur la base de coefficients MFCC [183], après analyse de leur saillance. Celle-ci est calculée à partir d'une collection de caractéristiques intrinsèques au signal acoustique telles que l'intensité, le contraste temporel, le contraste spectral et l'orientation. L'intérêt de leur approche est que l'analyse de la saillance est effectuée sur des signaux complexes [32] afin de réduire la dimension des données audio à traiter par le système de classification classique employé. Ainsi, et grâce à l'utilisation d'un algorithme de type LPS (adaptation des algorithmes de type *Latent Semantic Indexing*, LSI utilisés dans l'analyse de documents textuels [184]), les auteurs parviennent à une réduction de près de 74% des données à traiter. De plus, ils sont capables de détecter dans des scènes complexes et réalistes des sons qui ne sont pas au premier plan acoustique, du point de vue de leur intensité. Cet algorithme aboutit à une légère amélioration de l'analyse de scènes audio par rapport à un algorithme utilisant toutes les données, i.e. sans analyse de leur saillance et donc sans réduction de la dimensionnalité. Leur algorithme a, de plus, l'avantage de traiter des données non structurées et d'une façon non-supervisée.

2.3.2.3 Saillance multimodale

Les modèles présentés précédemment traitent la saillance audio ou la saillance visuelle séparément. Comme nous l'avons montré à la **Sec. 2.2.4** et la **Sec. 2.3.1.3**, l'intégration de données multimodales est un mécanisme essentiel à l'analyse performante d'un environnement, du point de vue perceptuel. De plus, la représentation multimodale d'entités perceptuelles permet également l'émergence de comportements attentionnels basés sur la congruence d'une modalité en fonction d'une autre. Le modèle présenté ici est une tentative de modélisation multimodale de la saillance.

En 2008, Jonas Ruesch *et al.* [185] ont développé un puissant modèle de filtrage attentionnel basé sur la saillance de données multimodales, implémenté dans le robot iCub [186, 187]. Leur algorithme, basé sur une *Sensory EgoSphere* [188] — carte interne au robot servant de représentation de la saillance multimodale — interprète les données audio et visuelles issues des capteurs du robot sous le prisme de leur saillance. La saillance visuelle, de façon similaire aux cartes décrites à la **Sec. 2.3.2.1**, est caractérisée par quatre sous-filtres de l'image analysée : intensité, teinte, direction, mouvement. La saillance audio est caractérisée, quant à elle, seule-

32. Base de données « The BBC Sound Effects Library Original Series », `http://www.sound-ideas.com`

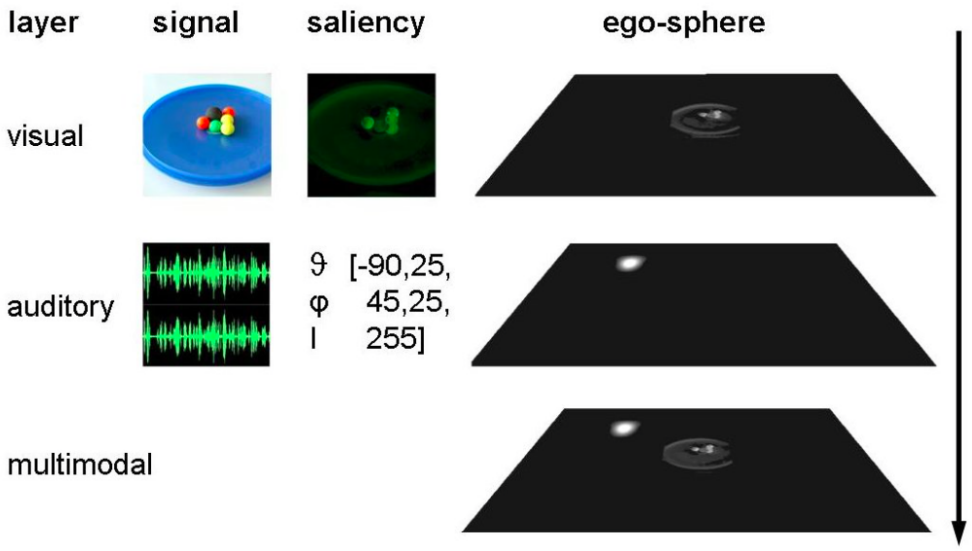


FIGURE 2.22 – CARTES DE SAILLANCE MULTIMODALES — Exemple d'agrégation de plusieurs cartes de saillance, d'après [185]. (*gauche* à *droite*) : saillance calculée à partir des signaux perçus par le robot iCub. (*haut* en *bas*) agrégation des différentes mesures de saillance et projection sur la carte égocentrique (concept de *Sensory Ego-Sphere*)

---

ment par les calculs de l'ITD [33] et ISD [34] de la trame audio courante. La **Fig. 2.22** illustre les cartes de saillance audio et visuelle ainsi que leur combinaison.

A cette carte, les auteurs y ont ajouté une autre carte d'*Inhibition* (*Inhibition Map*) pondérant la carte de saillance créée précédemment dans le but de limiter temporellement l'attraction pour un point de l'espace défini comme saillant et favoriser ainsi l'exploration de nouveaux points. Cette carte supplémentaire est construite selon une fonction de pondération gaussienne. D'autre part, de nombreux seuils et paramètres définis au préalable sont nécessaires pour que l'algorithme fonctionne au mieux. Bien que cet algorithme donne au robot la capacité de détecter un objet saillant — et donc éventuellement important — dans un environnement restreint, il ne prend pas en compte le contexte dans lequel cet objet se situe. Selon RUESCH *et al.* [185] :

« *A talking face is usually considered more* salient *than a silent one* [35]. »

Cette assertion suppose que certains stimuli sont, par nature, plus saillants que d'autres. Cependant, si différentes occurrences d'un stimulus considéré comme saillant apparaissent à différents intervalles, il sera toujours considéré comme saillant. Par exemple, si plusieurs personnes parlent, elles seront toutes considérées comme saillantes et aucun phénomène d'habituation ne va être pris en compte. Bien que la *carte de saillance égocentrique* corresponde à une mémoire à court-terme, d'après les auteurs, cette mémoire n'est utilisée que pour éventuellement réexplorer des régions

33. *Interaural Time Difference* — Différence temporelle interaurale, cf. **Sec. 2.2.1**.
34. *Interaural Spectral Difference* — Différence spectrale interaurale, cf. **Sec. 2.2.1**.
35. « Un visage qui parle est habituellement considéré comme plus *saillant* qu'un visage silencieux. »

déjà connues et non pour mettre à jour le niveau de saillance des objets audiovisuels déjà détectés. La saillance définie ici n'est donc pas un critère évoluant dans le temps et ne bénéficie pas d'apprentissage permettant d'intégrer un stimulus au sein d'un ensemble plus large d'événements audiovisuels.

2.3.2.4 Discussion

Nous avons présenté ici différents modèles de saillance audio, visuelle et audiovisuelle en tant que prémisses de l'élaboration d'un comportement attentionnel haut-niveau. Paraphrasant Kalinli *et al.* [180] concernant leur carte de saillance audio : cette analyse de la scène audio permet de déterminer quelle entité auditive nécessite éventuellement une attention particulière et donc une réquisition particulière des centres analytiques, *but de tout système attentionnel*. Les modèles audio décrits permettent de détecter des caractéristiques acoustiques d'intérêt au sein d'environnements sonores divers. Du point de vue visuel, les modèles de Itti *et al.* et d'Oliva *et al.* présentés ici montrent que l'utilisation de la saillance, qu'elle soit seule ou combinée à des données contextuelles, permet de faire émerger un comportement attentionnel aboutissant à une exploration visuelle proche de celle observée chez l'humain.

Selon Kayser *et al.*, il est possible de formuler deux hypothèses sur le principe de saillance perceptuelle, qu'elle soit audio ou visuelle :

> « *Either saliency is extracted by similar mechanisms implemented in both pathways*[36]*, or saliency for both systems*[37] *is extracted by the same multimodal cortical areas*[38]*.* »

Ici encore, l'intégration des informations multimodales apparaît comme un mécanisme au centre de processus complexes et aux rôles prépondérants que ce soit dans la représentation efficace et riche d'un environnement que dans l'analyse de la saillance d'un événement.

2.3.3 Conclusion

Dans cette section dédiée à l'*Attention* ont été successivement présentés les principaux mécanismes cérébraux impliqués dans la réaction motrice aux événements apparaissant dans un environnement et l'importance de l'intégration multimodale des informations sensorielles, notamment visuelles et auditives. Puis des concepts plus précis comme la *saillance* d'un événement et la réponse neuronale causée par un événement incongru (la MMN) ont été détaillés. Enfin, différents modèles attentionnels computationnels, parfois intégrés dans des plateformes robotiques, basés sur la Saillance ont été présentés.

36. Voies auditive et visuelle.
37. Systèmes auditif et visuel.
38. « Soit la saillance est extraite selon des mécanismes similaires dans les deux voies (auditive et visuelle), soit la saillance des deux systèmes (auditif et visuel) est extraire par les mêmes aires corticales multimodales. »

Nous retenons en particulier le phénomène de « *Mismatch Negativity* » et de *Saillance* en tant que bases conceptuelles fortes du modèle HTM. La MMN, d'un côté, réaction à un stimulus imprédictible apparaissant très tôt dans les aires cérébrales sensorielles, montre que la perception des stimuli est sous l'influence d'une capacité de prédiction des aires sensorielles et ce, afin d'accélérer le processus d'analyse de ces stimuli. Cela fait directement écho à la Théorie de la Hiérarchie Inverse (cf. **Sec. 2.2.3**) en cela que lorsqu'un stimulus devient prédictible — et cela arrive très rapidement : seulement quelques répétitions de ce stimulus suffisent — son analyse est alors anticipée. En revanche, lorsqu'un stimulus déviant apparait, il est nécessaire de réaliser à nouveau une analyse complète afin de retrouver les caractéristiques bas-niveau des signaux qui ont été délaissées.

De l'autre côté, la *Saillance* est une caractéristique fondamentale d'une entité audio ou visuelle au sein d'un contexte donné. De nombreux algorithmes développés au sein de plateformes robotiques permettent de faire émerger des comportements réactionnels performants, notamment dans le cadre de la modélisation des mouvements des yeux de type saccades oculaires. La saillance d'une entité est, la plupart du temps, définie par les caractéristiques intrinsèques du signal audio ou visuel perçu par les capteurs du robot : intensité, contraste, fréquence, temporalité etc. De plus, la plupart des modèles décrits plus haut n'intègrent que peu d'apprentissage à long-terme en cela que les algorithmes ne sont souvent que testés dans un type d'environnement seulement. Dans le cadre du modèle HTM, nous nous plaçons (i) sur une échelle temporelle plus longue et (ii) à un niveau plus cognitif, éloigné des caractéristiques bas-niveau des signaux. Ainsi, nous considérons qu'il est possible de faire émerger un comportement attentionnel pertinent en se plaçant au niveau sémantique de l'analyse des informations sensorielles. Au sein du modèle, les principes de MMN et de Saillance ont appuyé la modélisation de la Congruence d'un événement audiovisuel en fonction de l'environnement dans lequel il se trouve, en cela que nous avons considéré que l'apparition d'un objet audiovisuel peut être assimilé à un stimulus répondant aux mêmes lois qui régissent la perception de stimuli plus simples (comme ceux présentés plus haut).

Le modèle HTM, du point de vue attentionnel, pourrait être une forme de modélisation de l'activité du colliculus supérieur (CS), en tant que système recevant des informations audio et visuelles, les intégrant et générant consécutivement une commande motrice. Cependant, là où le CS traite des informations bas-niveau (il reçoit, par exemple, les informations visuelles avant qu'elles ne soient traitées par le cortex visuel primaire), notre modèle traite des données au niveau sémantique. D'autre part, le modèle HTM se place dans le cadre de l'attention exogène et endogène. La réaction rapide formalisée par les mouvements de tête lorsqu'un événement imprédictible survient est un processus attentionnel *exogène* car causé par un événement extérieur au robot et non par sa propre volonté. En revanche, elle est suivie par une forme d'attention *endogène* en cela que le système va ensuite décider de porter son attention sur l'entité audiovisuelle ayant causé le mouvement de tête préalable. Ces caractéristiques rappellent les travaux de Corbetta *et al.* sur la réorientation attentionnelle et les structures cérébrales qui permettent l'allocation pertinente et puissante des ressources cérébrales nécessaires à l'accomplissement d'une tâche donnée.

Le modèle HTM, en plus de modéliser un comportement réactif basé sur la pré-

dictibilité d'un événement, étant donné l'environnement dans lequel il se situe, est doté d'un mécanisme d'apprentissage en ligne des informations perçues par le robot. Cet apprentissage est indispensable à l'élaboration d'une représentation interne de l'environnement adaptée, robuste et pertinente. Ainsi, la section suivante décrit les principaux paradigmes d'apprentissage utilisés dans la communauté et détaillera en particulier le principe des cartes auto-organisatrices, base sur laquelle la partie « apprentissage » du modèle HTM a été créée.

2.4 Apprentissage

Q. Quel paradigme d'apprentissage choisir afin qu'un robot puisse apprendre la relation entre les entités audiovisuelles pendant l'exploration d'un environnement inconnu constitué de données non prédictibles et sans accès préalable à des informations sur cet environnement ?

Une des ambitions du modèle HTM est de doter le robot de la faculté d'apprendre de son exploration de l'environnement afin de pouvoir créer une representation stable et pertinente de celui-ci. Afin de déterminer quelle approche, quel paradigme, quel algorithme utiliser, il est nécessaire de répondre à plusieurs questions. Mais avant d'y répondre, il est ici nécessaire de décrire, ou rappeler, succintement certaines caractéristiques du modèle HTM, afin de comprendre le but précis de cet apprentissage ainsi que les contraintes l'encadrant.

Au sein du modèle HTM, un environnement est défini comme *l'ensemble des entités audiovisuelles qui le composent*. Ainsi, la carte cognitive que le modèle va tenter de créer sera composée de ces entités. La réaction attentionnelle générée par les entités audiovisuelles présentes dans l'environnement sera formalisée par les rotations de la tête du robot. Ces rotations ont une portée limitée puisqu'inspirée par les rotations possibles par une tête humaine : la tête ne peut tourner que de 90° à gauche ou à droite. Ainsi, dans le cas où un mouvement de tête n'est pas suffisant pour porter l'entité audiovisuelle dans le champ de vision, un mouvement de la base mobile sur laquelle le torse et la tête sont montées (la description complète du robot se fera à la **Sec. 3.3**) sera nécessaire. Mais sachant que (i) le but du modèle est autant de pouvoir déclencher des mouvements de tête vers des entités présentant un intérêt que de les inhiber, si l'entité n'est pas d'intérêt, et (ii) que la notion d'objet définie au sein du modèle HTM est multimodale (audio et visuelle), il est nécessaire d'être capable, lorsqu'une entité audiovisuelle est située derrière le robot de pouvoir accéder à l'information visuelle manquante. Cette information visuelle manquante, si elle peut être retrouvée à partir de la catégorie audio perçue, permettra ainsi de retrouver une représentation sous forme d'objet audiovisuel et ainsi de pouvoir poursuivre l'analyse effectuée par le modèle HTM, notamment celle basée sur la Congruence de cette entité au sein de cet environnement. L'apprentissage consiste donc ici en parvenir à modéliser le lien qui existe entre la modalité audio et la modalité visuelle.

De très nombreux algorithmes et paradigmes d'apprentissage ont été développés depuis les soixante dernières années. Avec l'augmentation considérable des capacités de calcul couplée à une forte baisse du prix des processeurs d'une part, et l'accès de plus en plus facile à des volumes de données impressionnants, le domaine de l'apprentissage machine a connu un bel essor. Ainsi, afin de sélectionner un algorithme d'apprentissage répondant aux besoins du modèle HTM, nous proposons à la section suivante de poser quatre questions fondamentales dont les réponses permettront d'identifier précisément les contraintes qui motiveront notre choix. Cette section a sa place dans ce chapitre d'état de l'art en cela que les réponses aux questions que nous allons poser nous permettront d'aborder les grandes catégories au sein desquelles les techniques d'apprentissage machine sont généralement classés.

2.4.1 Critères de décision de l'algorithme d'apprentissage

(a) le robot a-t-il accès à des informations sur l'environnement avant de commencer l'exploration ? Durant toute la conception du modèle HTM, nous nous sommes placés dans le cas d'un environnement totalement inconnu. Le robot, lorsqu'il pénètre cet environnement est complètement naïf. Attention cependant, il possède déjà certaines connaissances apprises avant son exploration : localisation et identification de sources sonores et visuelles notamment (qui seront détaillés au **Chap. 3**). En revanche, aucune connaissance ne lui a été donnée sur le *contenu* de l'environnement en terme d'objets audiovisuels. Par extension, aucune règle de *comportement* non plus n'est donnée au robot : il s'agit justement d'un environnement inconnu et le robot doit déterminer lui-même le comportement attentionnel à adopter en fonction de cet environnement et des objets qui y sont présents. Sachant cela, il est possible de choisir entre les deux paradigmes d'apprentissage suivant, conditionnant le choix futur d'un algorithme précis :

- *hors-ligne* (*offline*) : acquérir **préalablement** le plus d'information possible sur le monde (enregistrements, scans, mesures etc), afin d'en tirer une représentation la plus exhaustive et précise possible. Toutes ces données *a priori* serviront alors à élaborer des modèles qui seront ensuite inculqués au robot avant même qu'il évolue dans ces environnements. Une fois qu'il sera dans le monde réel, il cherchera à retrouver quel modèle appris préalablement ressemble le plus à celui qu'il est en train d'explorer. A cet environnement sera associé un ensemble de règles de comportement qu'il appliquera alors. L'idée est donc de connaître suffisamment bien le monde pour pouvoir le *modéliser* presque parfaitement.

- *en ligne* (*online*) : acquérir également le plus possible d'information mais **lors de l'exploration** de l'environnement. Le robot est donc possiblement naïf au moment où il entre dans un environnement. Le but est que le robot puisse créer sa propre représentation interne du monde qu'il découvre, et non la représentation de l'expérimentateur.

Cette distinction entre paradigmes hors-ligne et en ligne ne peut pas être directement opposée au fait que le robot doive apprendre d'un environnement inconnu. Il est tout à fait possible d'utiliser un paradigme d'apprentissage en ligne lors de l'exploration tout ayant effectué une étape d'apprentissage préalable permettant de faciliter l'analyse ultérieure des informations captées lors de cette exploration. Mais dans le cas du modèle HTM, l'apprentissage de données hors-ligne n'est pas possible : nous ne pouvons pas savoir quelles catégories audiovisuelles seront présentes dans les environnements explorés. Il serait malgré tout possible de prédéfinir toutes les combinaisons audiovisuelles possibles, en fonction des classes audio et visuelles à disposition. Cette option a comme limite une incapacité relative à être généralisée : l'ajout de nouveaux experts d'identification entraînerait la redéfinition de tous les appariements possibles, de même que l'ajout d'une nouvelle modalité (données tactiles par exemple). Une autre limite de cette approche est l'augmentation significative de la dimension de l'espace audiovisuel ainsi défini. Cet espace contient, en outre, beaucoup d'informations non pertinentes : l'ensemble des combinaisons possibles est très grand face aux combinaisons auxquelles le robot va réellement être confronté. Il

aurait également été possible d'indiquer clairement quelles sont les catégories audio-visuelles « réalistes » que le robot pourra rencontrer lors de ses explorations. Mais ici survient le problème de la connaissance inculquée au système par l'expérimentateur, connaissance qui n'est donc pas réellement apprise puisqu'elle simplement transférée. De plus, cela sous-entend qu'il existe une « vérité » à apprendre, notion profondément subjective. Nous préférons que le robot construise sa propre subjectivité.

Se baser sur un paradigme d'apprentissage en ligne permettra de laisser le robot apprendre lui-même des données qu'il perçoit. A partir de ces données, il pourra créer sa représentation interne de l'environnement par les objets audiovisuels qui y sont présents. Le choix d'un paradigme de type en ligne a conditionné notre sélection d'un algorithme d'apprentissage en cela qu'il a un impact sur l'aspect temporel (l'algorithme n'a pas le même temps dont dispose ceux hors-ligne) et l'aspect dimensionnel (l'algorithme ne doit pas nécessiter un grand nombre de données pour converger).

(b) Connaît-on l'état final vers lequel le système doit converger ? Autrement dit, est-il possible de savoir si le réseau apprend correctement ?

Oui et non.

Oui, car afin de valider le modèle HTM, nous avons développé des critères d'évaluation de ses performances tant au niveau de l'apprentissage des règles de Congruence que de celui du lien multimodal existant entre une catégorie audio et une catégorie visuelle. Non, car l'algorithme d'apprentissage n'a pas accès à cet état final. Les trois grands paradigmes d'apprentissage que nous pouvons considérer ici sont : supervisé, non-supervisé et auto-supervisé / par renforcement.

L'apprentissage **supervisé** consiste en faire converger un système vers un état final connu à l'avance [189]. Par exemple, l'apprentissage des experts de localisation audio utilisés au sein du projet Two!Ears est supervisé, car le système doit apprendre à lier les données audio perçues (après extraction de caractéristiques des signaux, cf. **Chap. 3**) à des positions en azimut connues. Ce genre de paradigme fonctionne extrêmement bien mais nécessite (i) de connaître les données à l'avance, (ii) d'effectuer une longue étape d'apprentissage avant que le robot ait un comportement autonome, et (iii) un très grand nombre de données : dans le cadre de la détection et la reconnaissance d'objets par exemple, les banques de données vont de cinq cents exemples (base de données de Berkeley BSDS500 [190]) à près de quinze millions (base de données ImageNet [191, 192]). D'ailleurs, la réduction de la dimension des données à apprendre constitue un enjeu majeur des algorithmes d'apprentissage supervisés tant certains d'entre eux peuvent nécessiter un grand volume de données [193].

L'apprentissage **non-supervisé** est inverse : l'état final n'est pas connu à l'avance et le système cherche à converger vers une solution permettant de catégoriser les données d'entrée en groupes distincts, compréhensibles et dont les caractéristiques sont utilisables pour des tâches ultérieures. Par exemple, dans le domaine du traitement des signaux audio, la détermination du bruit de fond dans le but de sa soustraction (puisque pouvant interférer avec les signaux d'intérêt) peut être soit effectué à par-

tir de modèles de bruits de fond et de signaux d'intérêt appris selon une approche supervisée [194, 195, 196], soit en utilisant des algorithmes non-supervisés justement et ne requérant donc pas de connaître au préalable les caractéristiques du signal à traiter : l'algorithme agit directement dessus, sans biais ni règle [197, 198, 199, 200].

Enfin, l'apprentissage **auto-supervisé** constitue en quelque sorte un mélange des paradigmes précédents [201]. Particulièrement adapté au contexte robotique, il est également une modélisation convaincante des processus d'apprentissage observés chez l'humain [202]. Par exemple, un débutant démarrant la pratique du piano va principalement se baser sur sa vue. Ensuite, il apprendra à lier sa perception visuelle à la sensibilité de ses doigts ainsi que leur position spatiale, dans le but d'accomplir la même tâche. *L'apprentissage du comportement sensorimoteur aura donc été supervisé par la vue.* L'apprentissage **auto-supervisé** entre dans la catégorie de l'apprentissage par renforcement [203] en cela que l'apprentissage d'un comportement grâce à un processus de type *trial-and-error* est compatible avec l'auto-supervision.

Parmi ces trois paradigmes, deux satisferaient nos contraintes : l'apprentissage non-supervisé et l'apprentissage auto-supervisé. Cependant, une donnée supplémentaire doit ici être mentionnée : la plateforme robotique utilisée dispose de mouvements de tête. Le modèle HTM ayant pour but d'apprendre la relation entre modalité audio et modalité visuelle, le système va avoir besoin d'accéder aux deux sources d'information afin de réaliser son apprentissage. Ainsi, des mouvements de tête vont être générés pour accéder aux informations visuelles, lorsque celles-ci ne sont pas disponibles, c'est-à-dire lorsque l'objet multimodal est situé en-dehors du champ de vision du robot. D'un autre côté, lorsque le système estimera qu'il aura suffisamment bien appris, il inhibera ces mouvements de tête, jugeant qu'il n'est plus nécessaire d'acquérir une information supplémentaire. Ce comportement se rapproche donc plutôt des algorithmes auto-supervisés en cela que le robot va lui-même juger les performances de son apprentissage et éventuellement l'arrêter lorsque ce n'est plus nécessaire. Ici apparait donc une considération temporelle : le moment où le système jugera que son apprentissage est suffisant dépendra du nombre de données auxquelles il a eu accès et du temps qu'il aura passé à les apprendre. Ainsi :

(c) Quel est le délai entre le moment où les données sont perçues et le moment où le robot doit réagir en fonction de cet apprentissage ? Avec l'adaptabilité offerte par les algorithmes d'apprentissage en ligne et la possibilité de ne pas connaître l'état final du système à atteindre dans le cas des algorithmes non/auto-supervisés, émerge le souci du délai potentiel entre le moment où le robot commence son exploration et celui où il devient capable de réagir avec cohérence (pour l'expérimentateur) aux événements se déroulant dans l'environnement, c'est-à-dire le moment où son apprentissage aura convergé vers un état robuste et pertinent. Deux paramètres temporels sont ici à considérer : (i) le moment où le robot perçoit la première donnée et le moment où il est capable d'émettre une première hypothèse viable sur l'environnement, et (ii) le temps entre la perception d'une donnée et son traitement consécutif. Le premier point concerne le temps de convergence du réseau, tandis que le second concerne la complexité computationnelle de l'algorithme utilisé. Ainsi, concernant la complexité computationnelle, un système d'apprentissage de type réseau de neurones artificiel est qualifié de « temps réel » s'il est capable de terminer une étape d'apprentissage d'une nouvelle application dans un temps

court face aux contraintes externes [204]. Ce temps peut être de l'ordre de la microseconde, de la milliseconde ou de la seconde, en fonction du problème à traiter. Ici, nous attendons un temps de traitement inférieur à la demi-seconde : comme exposé au **Chap. 3**, une trame audio ou visuelle dure 500 ms. Ainsi, afin de ne pas causer un retard dans le traitement des données aboutissant soit à la non acquisition de nouvelles données soit au décalage entre les données analysées et les données réelles, nous devons utiliser un algorithme suffisamment puissant pour qu'une itération d'apprentissage soit la plus rapide possible. D'autre part, concernant le temps de convergence, le système a pour ambition de permettre au robot de prendre des décisions quant à ses mouvements de tête très rapidement après le début de son exploration. Ainsi, nous devons utiliser un algorithme au temps de convergence rapide. Ce temps est fonction de l'architecture de l'algorithme mais également des données à traiter : plus les données sont complexes, variées et nombreuses, plus le système d'apprentissage mettra du temps à converger. Ce qui nous amène à notre quatrième question :

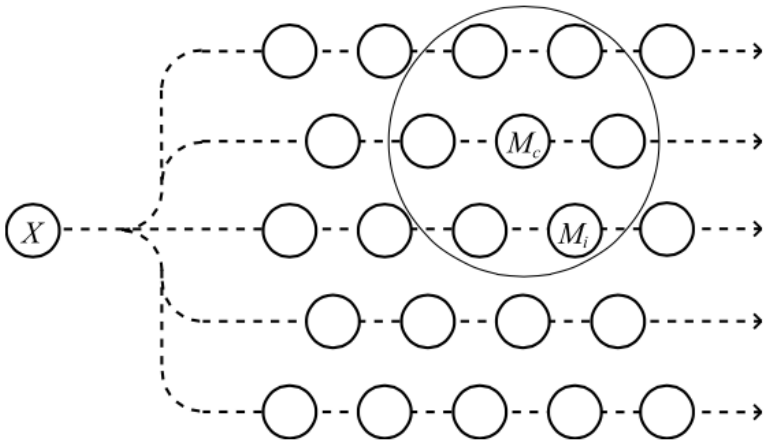
(d) Quelle est la complexité des données à traiter ? Il s'agit peut-être de l'élément le plus important dans la sélection d'un système d'apprentissage. En effet, chaque système a ses points forts et ses points faibles en fonction de la façon dont les données sont faites : type de données, présence de redondance, présence de valeurs nulles, taille des données, variabilité, données organisées en catégories ou non etc. Dans notre cas, les données seront des *vecteurs de probabilités* d'appartenance à une catégorie audio ou visuelle. Dans les cas les plus complexes que nous avons traités en environnements simulés, nous avons eu des vecteurs ayant jusqu'à 50 composantes. Cela représente des données de plutôt faible dimension et complexité.

2.4.1.1 Discussion

D'après l'ensemble des réponses apportées aux questions posées ci-dessus, voilà la liste des contraintes que l'algorithme d'apprentissage doit respecter :

1. apprentissage non-supervisé ou auto-supervisé : pas d'état final auquel le système peut se référer pour guider son apprentissage ;
2. faible nombre de données : le robot doit pouvoir apprendre à partir d'un jeu de données très restreint puisque ces données arrivent une par une au cours de l'exploration et selon une dynamique assez lente (toutes les 500 ms) ;
3. rapide (quasi temps réel) : le robot doit pouvoir réagir rapidement aux nouvelles informations acquises lors de son exploration puisqu'il est censé pouvoir générer une réponse motrice comportementale le plus rapidement possible sur la base des connaissances apprises ;
4. robustesse : le résultat de l'apprentissage conditionnant la réaction du robot et les données étant issues d'experts faisant parfois des erreurs, il est également nécessaire que l'algorithme puisse être suffisamment robuste pour prendre en compte cette variabilité.

Après avoir étudié de nombreux algorithmes d'apprentissage, celui qui a le mieux rempli les conditions d'utilisation du modèle HTM est l'algorithme de **carte auto-adaptatrice** [205]. En effet, ce type d'algorithme est non-supervisé, peut apprendre



Carte Auto-Organisatrice —

Figure 2.23 – Illustration d'une carte auto-organisatrice (*Self-Organizing Map*). Un vecteur d'entrée est envoyé à un ensemble de modèles (ou nœuds) M_i. Parmi tous ces modèles, M_c est celui qui représente le mieux la donnée d'entrée. Une étape d'apprentissage correspond à étendre cette similarité entre M_i et M_c à tous les modèles présents dans le voisinage direct de M_c (figure d'après [206]).

avec peu de données, est rapide à exécuter et son architecture permet de le manipuler et de le modifier très facilement. De plus, il a été énormément étudié et utilisé depuis les trente dernières années (cf. section suivante), les théorèmes de convergence sont donc bien connus, tout comme ses limitations. D'autre part, étant donné le contexte dans lequel cet apprentissage va avoir lieu, détaillé au **Chap. 6**, notre approche, utilisant une carte auto-adaptatrice, se place également dans la catégorie de l'apprentissage par renforcement.

Une carte auto-adaptatrice, ou carte auto-organisatrice, carte de Kohonen, ou encore *Self-Organizing Map* (SOM), est un réseau de neurones artificiel permettant d'avoir une représentation discrète, le plus souvent en deux dimensions, d'un espace d'entrée selon une approche non-supervisée [205, 207, 208]. Le but principal est d'effectuer une réduction de la dimension des données d'entrée *via* une forme de quantification vectorielle de celles-ci selon une méthode de construction de carte itérative et auto organisée (cf. **Fig. 2.23**). L'algorithme a été développé au début des années 80 par Teuvo Kohnonen et est devenu extrêmement populaire dans la communauté de l'apprentissage puisqu'il offre une visualisation de données souvent très larges dans un espace réduit tout en préservant la topologie de l'espace d'entrée (cf. [209, 210, 211] pour une liste des quelque 7600 papiers utilisant l'algorithme des cartes auto-adaptatrices, de 1981 à 2005). De plus, le fait qu'il soit non-supervisé, donc qu'aucune donnée *a priori* n'ait besoin d'être connue et implémentée par l'expérimentateur pour que le réseau converge vers une représentation pertinente de l'espace d'entrée, ajoute à son intérêt et son succès. L'idée originale, inspirée du fonctionnement de certaines zones du cortex, est de représenter des données *multi-dimensionnelles* dans un espace *bidimensionnel*. Autrement dit, il s'agit de détecter des ressemblances dans la distribution des données d'entrée afin de les regrouper puis d'assigner à chacun des groupes déterminés une zone particulière de l'espace de représentation bidimensionnel. Ainsi, après apprentissage, chacune des zones de la

carte créées lors de l'apprentissage code une information similaire. Il est courant de parler de la *tonotopie* des cartes auto-organisatrices, de la même façon que l'on parle de la tonotopie du cortex auditif. La formalisation complète de ce type d'algorithme d'apprentissage sera effectuée au **Chap. 6**.

2.4.2 Fusion de classifieurs

La fusion de classifieurs est le pendant computationnel de l'intégration multimodale, traitée à la **Sec. 2.2.4**. Dans cette section seront donc seulement exposés les différents paradigmes de fusion de classifieurs existant. Les classifieurs sont des entités computationnelles analytiques permettant de passer d'une information brute (signaux audio ou image par exemple) à une interprétation de cette information. Cette interprétation vise le plus souvent à extraire des caractéristiques de l'information brute afin de lui assigner une catégorie. Le but des classifieurs est (i) de réduire la dimensionnalité des données d'entrées, (ii) de catégoriser ces données en fonction de critères variables, dépendant de l'étape d'extraction de caractéristiques et (iii) de passer d'une représentation brute à une représentation plus symbolique. Un système robotique doté de plusieurs capteurs, comme le robot sur lequel le système TWO!EARS a été intégré, peut posséder des classifieurs permettant de catégoriser les données perçues par chacun d'entre eux. Par exemple, dans le système TWO!EARS, les données audio sont traitées par une série de classifieurs permettant de leur assigner une probabilité d'appartenance à une catégorie donnée, de même pour la modalité visuelle. La fusion peut se faire entre des classifieurs de même nature, comme l'ensemble de ceux dédiés à l'identification audio, ou de nature différente comme l'ensemble des experts d'identification audio et l'ensemble des experts d'identification visuelle. Dans notre cas, la fusion sera double : dans un premier temps, une fusion **intra**modale va permettre de prendre une décision sur l'ensemble des classifieurs au sein d'une même modalité ; dans un second temps, une fusion **inter**modale va permettre de combiner les décisions prises au niveau intramodal afin de prendre une nouvelle décision sur l'appartenance d'un objet audiovisuel à une catégorie audiovisuelle. L'idée est donc de combiner des résultats obtenus sur l'audio, d'une part, et la vision, d'autre part, afin de faire émerger une nouvelle donnée : la notion d'*objet multimodal*.

La fusion de classifieurs est effectuée par des *systèmes de support décisionnel* (*Decisional Support Systems*, DSS). Selon DYMITR RUTA & BOGDAN GABRYS [212] :

> « *The objective of all decision support systems is to create a model, which given a minimum amount of input data/information, is able to produce correct decisions*[39]. »

La fusion de classifieurs a été particulièrement développée, utilisée et étudiée depuis une vingtaine d'années et est maintenant intensivement employée chaque fois que différentes sources d'informations sont disponibles pour caractériser une entité à analyser par un système intelligent [213, 214, 215, 216, 217] (et [212, 218, 219, 220, 221] pour différentes revues de la fusion multimodale computationnelle). De nombreuses

39. « L'objectif de tout système de support décisionnel est de créer un modèle qui, avec un minimum de données / d'information, est capable de prendre des décisions correctes. »

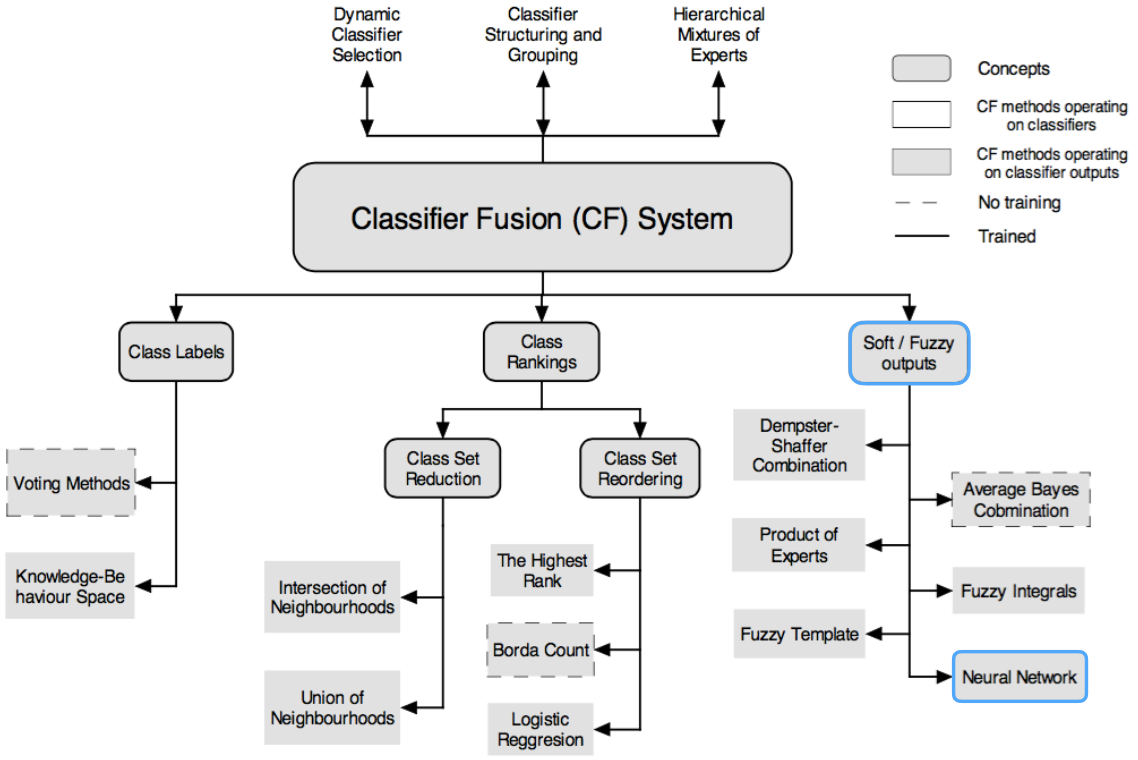


FIGURE 2.24 – PARADIGMES DE FUSION DE CLASSIFIEURS — Le paradigme de fusion utilisé par le modèle HTM se place dans la catégorie *Soft/Fuzzy Outputs — Neural Networks* (rectangles bleus) (figure d'après [212]).

---

stratégies existent pour la fusion de classifieurs. Mais au préalable, et de façon similaire au choix d'un algorithme d'apprentissage, des questions sont à se poser avant d'effectuer une fusion pertinente des données auxquelles le système a accès [221].

(a) A quel niveau fusionner ? *Au niveau des caractéristiques bas-niveau du signal (*features level*) ? ou au niveau de leur interprétation plus haut niveau (*semantic level [222, 223]*) ?* Nous effectuerons une fusion au niveau **sémantique**. Les caractéristiques bas-niveau des signaux comme le contenu spectral ou la dynamique temporelle (« *onset & offset time* » pour les signaux de parole par exemple) ne seront pas pris en compte : le modèle HTM se situe en sortie des classifieurs.

(b) Comment effectuer cette fusion ? *Quelle stratégie et quel algorithme utiliser en regard des caractéristiques des données à disposition et du problème à résoudre ?* Une fusion basée sur un type de quantification vectorielle des données d'entrée sera utilisée. Elle sera détaillée dans la partie consacrée au *Multimodal Self-Organizing Map* (cf. **Sec. 6.2**).

(c) Quand effectuer la fusion ? *Toutes les données arrivent-elles au même moment ? Ont-elles la même dynamique temporelle ?* Les données auxquelles le modèle HTM aura accès seront synchronisées. Ainsi, les données visuelles et audio sont

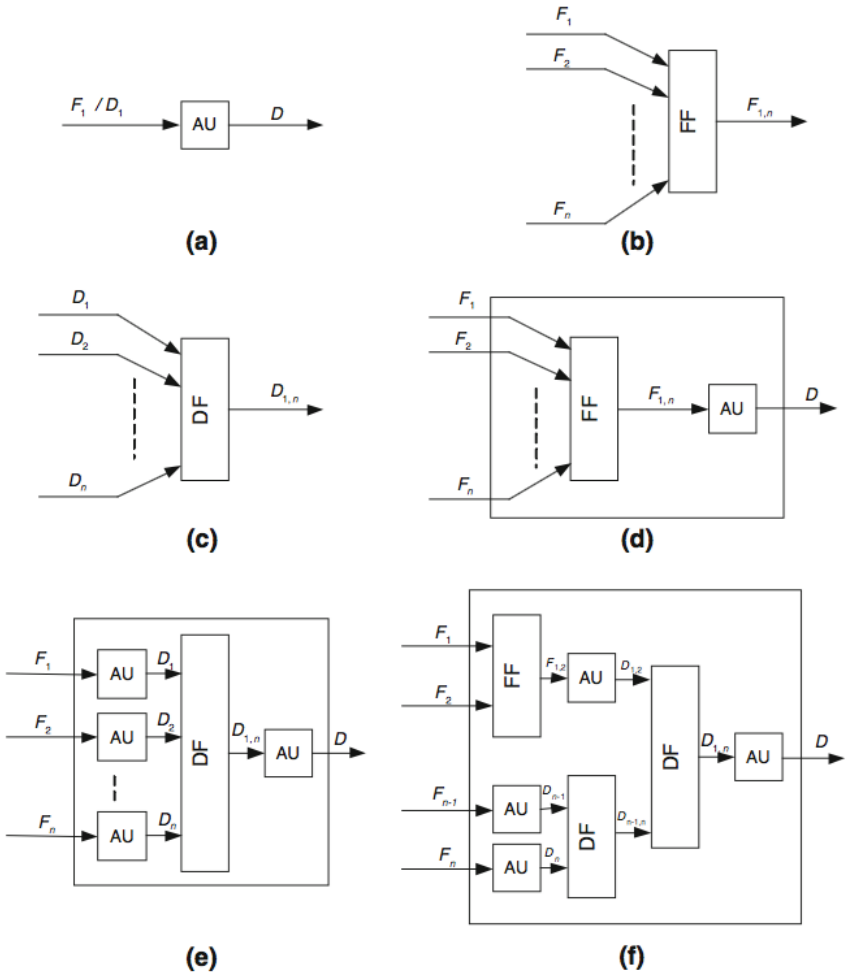


FIGURE 2.25 – STRATÉGIES DE FUSION MULTIMODALE — (*a*) *Analysis unit*, (*b*) *feature fusion unit*, (*c*) *decision fusion unit*, (*d*) *feature level multimodal analysis*, (*e*) *decision level multimodal analysis*, (*f*) *hybrid multimodal analysis* (figure d'après [221]).

disponibles au même moment, qu'il s'agisse des résultats de localisation ou d'identification. D'autre part, l'ambition du modèle HTM, est de donner au robot la capacité de réagir très rapidement aux événements apparaissant dans l'environnement. Ainsi, la fusion se fera à l'échelle temporelle de la trame afin de fournir au robot, de façon synchrone, une estimation de la catégorie audiovisuelle à laquelle l'objet considéré appartient.

(d) Quelles données fusionner ? *Toutes les données sont-elles pertinentes dans le processus de fusion ?* Dans un premier temps, les données issues des experts d'identification visuels et audio seront fusionnées. Une extension aux données de localisation audio et visuelle sera proposée en fin de manuscrit.

2.4.2.1 Discussion

La **Fig. 2.24** illustre les différents types de fusion de classifieurs existants, chacun étant dédié à un type de fusion et à un type de données. Dans notre cas, nous

récupérerons les sorties des classifieurs et nous plaçons donc au niveau de la branche « *Soft/Fuzzy outputs* » du diagramme. Chaque sortie des classifieurs est une probabilité d'appartenance à la classe pour laquelle le classifieur a été entraîné et est compris entre $[0, 1]$. Ces sorties sont appelées *mesures floues* [224] (« *fuzzy measures* ») en cela qu'elles décrivent différentes dimensions de l'incertitude de l'information qu'elles traitent (notamment les signaux audio, le comportement des classifieurs visuels étant légèrement différent, comme détaillé à la **Sec. 3.2.3**). Le but de la fusion de ce type de classifieurs est donc de réduire le niveau d'incertitude en maximisant l'apport des contributions de chaque classifieurs. Le paradigme d'apprentissage que nous avons choisi à la section précédente pour l'apprentissage des données audiovisuelles effectuera une fusion des classifieurs, comme détaillé à la **Sec. 6.2**. Ainsi, nous nous plaçons dans le cas de la fusion par réseau de neurones artificiels (« *Neural Network* » dans le diagramme). La formalisation de cette fusion sera effectuée à la **Sec. 6.3**.

D'autre part, la **Fig. 2.25** schématise les différents types de stratégies de fusion multimodale et particulièrement le niveau auquel cette fusion est effectuée : directement au niveau des caractéristiques des signaux ou en aval de ceux-ci, après que des éléments décisionnels aient réduit l'espace des données initiales en les transformant en « catégories » ou « groupes ». Nous nous plaçons dans le cas du schéma (e) : les données extraites des signaux sont traitées par des classifieurs (« *analysis units* », AU) qui émettent une décision D de type « *fuzzy output* » (probabilité d'appartenance). L'ensemble des décisions sont ensuite rassemblées par le modèle HTM (correspondant à D_f) et une décision unique est prise par une nouvelle AU. Cette décision correspondra séparément à la détermination de la classe audio et visuelle à chaque trame. La combinaison de deux de ces types de DSS permettra ensuite d'obtenir la catégorie audiovisuelle.

2.4.3 Conclusion

Cette section a été consacrée au très vaste domaine de l'apprentissage machine. Dans un premier temps, nous avons défini précisément le problème que nous cherchons à résoudre ainsi que les contraintes externes auxquelles le modèle HTM est soumis. Parmi celles-ci : la non-supervision de l'apprentissage, l'absence de données *a priori* et une complexité algorithmique relativement faible du fait de son embarquement dans un robot réel. Nous avons finalement convergé vers l'emploi d'une carte auto-organisatrice dont nous détaillerons la formalisation au **Chap. 6**. Cependant, nous verrons que, tel quel, l'algorithme SOM n'est pas entièrement capable de répondre aux besoins exprimés par le modèle HTM. Nous avons créé le *Multimodal*-SOM constituant une adaptation du SOM à nos contraintes, tels que nous l'avions proposé en 2016 [225]. Grâce au M-SOM, et conjointement avec toute l'architecture du modèle HTM, notre algorithme d'apprentissage sera en mesure d'apprendre le lien entre audio et vision, de corriger les éventuelles erreurs de classification des experts d'identification ainsi que d'effectuer une fusion de classifieurs. La formalisation du M-SOM sera faite à la **Sec. 6.2**.

Nous souhaitons par ailleurs rappeler ici le modèle de Kuniaki Noda *et al.* [121] que nous avons décrit à la **Sec. 2.2.4.2**, modèle d'intégration multimodale (audio, vision et données odométriques) porté sur une plateforme robotique humanoïde et assez

proche de ce que nous cherchons à développer au sein du modèle HTM. Leur modèle utilise un à trois réseaux de neurones profonds (DNN) afin d'effectuer l'intégration cross-modale, l'inférence de données manquantes et la prédiction de nouvelles données. Nous avons déjà expliqué auparavant les raisons pour lesquelles ce modèle ne répondait pas aux contraintes — matérielles ou conceptuelles — auxquelles nous sommes soumis. Nous avons choisi ici d'utiliser un système beaucoup plus simple, composé d'un réseau de neurones traditionnel. Grâce à l'adaptation de l'algorithme de SOM à notre problème, nous pensons également parvenir à une intégration multimodale de données, portée sur un robot et explorant de façon non supervisée son environnement.

2.5 Conclusion du Chapitre

Ce chapitre a passé en revue tous les concepts sur lesquels le modèle HTM se base : exploration, perception, processus attentionnels et apprentissage machine. Leurs fondements biologiques et particulièrement neuronaux ont également été exposés, permettant de saisir les inspirations de nombreuses applications computationnelles et robotiques cherchant à modéliser ces comportements observés chez l'animal, et l'homme en particulier. De tous ces travaux de recherche détaillés, nous allons essayer d'en faire une synthèse afin de mettre en avant tout ce qui a contribué à l'élaboration du modèle HTM. Dans toute cette fin de chapitre, nous parlerons du modèle HTM en tant que le système implémenté ainsi que son incarnation dans le robot mobile, robot décrit au chapitre suivant.

Tout d'abord, notre modèle sera *actif* lors d'une tâche d'exploration. Les bases neurales de l'exploration animale nous montrent que la représentation d'un environnement sous forme de cartes est communément admise et prouvée. De façon similaire, la communauté robotique a adopté ce type de représentation afin de rendre compte de l'état qualitatif et quantitatif de l'exploration d'un environnement. Cependant, la plupart des algorithmes d'exploration se concentrent sur l'acquisition d'informations topologiques sur l'environnement. Lorsque certains algorithmes prennent en compte des éléments plus sémantiques, comme des obstacles ou des entités non-statiques (des personnes), il s'agit de les intégrer dans une planification de navigation mais rarement comme une source d'interaction modulant possiblement l'exploration du robot.

Ensuite, le modèle HTM effectuera une intégration multimodale des informations perçues dans l'environnement, en particulier les informations des capteurs audio et visuels. La façon dont les structures cérébrales comme le colliculus supérieur ou le système vestibulaire reçoivent très rapidement, dans le transit de l'information, les données sensorielles — qu'elles soient directement issues des capteurs ou qu'elles viennent des aires sensorielles — montre que l'intégration multimodale est essentielle à une compréhension riche et robuste de l'environnement. De plus, ces aires intégratives ont également la capacité de générer des ordres moteurs en réaction aux informations sensorielles perçues. Ainsi, ces structures jouent un rôle prépondérant dans l'émergence des phénomènes attentionnels exogènes — provoqués par l'environnement et non par l'organisme. Le modèle HTM est également une modélisation de processus attentionnel, formalisé par des mouvements de tête vers des sources

audiovisuelles incongrues. En cela, il se rapproche beaucoup du fonctionnement du collicus supérieur et du système vestibulaire, deux structures pouvant générer des mouvements de tête.

D'autre part, le modèle HTM tente également d'intégrer les informations perçues sous forme de carte qui sera utilisée afin de sélectionner la cible des prochains mouvements de tête. Cette représentation nécessite l'apprentissage du lien cross-modal existant entre les informations visuelles et auditives. Parmi tous les paradigmes et techniques explorés, l'algorithme des cartes auto-organisatrices a été celui que nous avons retenu en raison de sa simplicité, son efficacité, son utilisation en ligne et fonctionnant selon un mode non-supervisé. Cependant, ce type de réseau de neurones artificiel n'étant pas compatible avec la gestion de données manquantes, nous avons dû le modifier en un *Multimodal-Self Organizing Map* (M-SOM) qui nous permettra d'inférer des données manquantes de façon rapide et robuste.

Ainsi, le modèle HTM peut être décrit globalement comme un système exploratoire et attentionnel bas-niveau permettant de créer une représentation interne d'un environnement inconnu à l'aide de mouvements de tête.

Chapitre 3

Two!Ears

Le modèle Head Turning Modulation a été développé au sein du projet européen Two!Ears, projet FET [1] qui a démarré le 1er décembre 2013 et s'est terminé le 30 novembre 2016, la revue finale ayant eu lieu les 19 & 20 janvier 2017. Le projet Two!Ears a pour but le développement d'un modèle computationnel intelligent et actif de la perception auditive binaurale en contexte multimodal. Le paradigme innovant sur lequel se base ce projet est de mêler une analyse ascendante des signaux audio, visuels ou proprioceptifs (approche *bottom-up*) avec une propagation descendante du résultat de cette analyse (approche *top-down*) dans le but de moduler le comportement du robot en fonction de la première. Par comportement, nous entendons aussi bien le caractère réactif du robot que la modulation dynamique des processus d'analyse des signaux perçus. Le système Two!Ears comporte une partie logicielle (modèles binauraux, algorithmes d'apprentissage, modèles attentionnels et réactifs etc.) et matérielle (robot mobile, capteurs visuels et auditifs). Le système Two!Ears est libre de droit et complètement *open-source*.

Le logiciel développé se base sur l'utilisation d'un système *Blackboard* [226] et d'un ensemble d'experts capables d'effectuer de nombreuses analyses. Le système élabore des hypothèses sur le monde en cours d'exploration, hypothèses à partir desquelles le robot va réagir en fonction des situations dans lesquelles il se trouve. L'audition binaurale est le véritable cœur de ce projet fortement bio-inspiré. En effet, Jens Blauert, entre autres, a montré au cours de sa carrière à quel point l'audition binaurale est un atout majeur des espèces animales dans la compréhension de leur environnement [227, 228]. Pour cela, l'intégration de processus descendants permet d'améliorer grandement l'analyse de scènes auditives en permettant l'intégration d'une forme de cognition du robot dans les processus de capture et d'analyse des signaux perçus. L'ambition globale du projet peut être résumée en cette phrase :

« *To read the world with two ears.* »

Une des applications phares du projet Two!Ears est l'exploration d'environnements complexes à des fins de recherche et de sauvetage de personnes [229] (*Search & Rescue*, S&R par la suite). Ce genre de scénario implique la capacité de se déplacer

1. *Future and Emergent Technologies* — Technologies Futures et Emergentes.

FIGURE 3.1 – CONSORTIUM TWO!EARS — Répartition des neuf laboratoires européens et du partenaire américain.

dans un environnement composé d'obstacles et comportant plusieurs sources audiovisuelles que le système doit pouvoir reconnaître rapidement afin de réagir en conséquence. Au-delà de l'aspect purement analytique de scènes audiovisuelles se situe l'aspect cognitif, avec notamment l'implémentation de processus attentionnels. En effet, dans un scénario S&R autant que dans n'importe quel contexte d'exploration « intelligente » par un robot humanoïde, la compréhension d'un environnement est indispensable à l'élaboration d'un comportement pertinent et adapté à la tâche à effectuer, qu'elle soit une tâche d'exploration, de sauvetage ou d'interaction avec les objets audiovisuels présents. Il est donc nécessaire pour le robot d'être capable d'assigner du *sens* aux objets qu'il perçoit et de pouvoir réagir en conséquence. La perception des signaux, leur analyse, l'exploration de l'environnement et les capacités réactives du robot sont toutes implémentées selon des paradigmes ayant l'ambition de donner au robot une autonomie complète : le robot doit pouvoir se passer de toute intervention humaine une fois qu'il est placé dans un environnement, connu ou non.

Neuf laboratoires européens et un partenaire américain composent le projet TWO!EARS :

Audio Visual Technology Group — Université Technologique d'Ilmenau (Allemagne)
ALEXANDER RAAKE (porteur du projet)
HAGEN WIERSTORF

Neural Information Processing Group — Université Technique de Berlin (Allemagne)
KLAUS OBERMAYER
IVO TROWITZSCH
JOHANNES MOHR
YOUSSEF KASHEF

Department of Electrical Engineering-Hearing Systems — Université Technique du Danemark
TORSTEN DAU
TOBIAS MAY

Institute of Communication Acoustics — Université de la Ruhr (Bochum, Allemagne)

Jens Blauert
Dorothea Kolossa
Thomas Walther
Cristopher Schymura

Institut des Systèmes Intelligents et de Robotique — Université Pierre et Marie Curie (Paris, France)
Bruno Gas
Sylvain Argentieri
Benjamin Cohen-Lhyver

Robotics, Action and Perception Group — Laboratoire d'Architecture et d'Analyse des Systèmes (Toulouse, France)
Patrick Danès
Ariel Podlubne
Thomas Forgue

Institute of Communications Engineering — Université de Rostock (Allemagne)
Sascha Spors
Fiete Wirstof

Department of Computer Science — Université de Sheffield (Grande-Bretagne)
Guy Brown
Ning Ma

Human-Technology Interaction Group — Université Technologique d'Eindhoven (Pays-Bas)
Armin Kohlrausch
Ryan Chungeun Kim

The Center for Cognition, Communication, and Culture Rensselaer (Etat de New-York, Etats-Unis)
Jonas Braasch

A chaque laboratoire a été assigné un ensemble de tâches à accomplir durant le projet et avec des contraintes de temps données. A l'instar de la plupart des projets européens, une organisation en groupes de travail (*Work Packages*, WP par la suite) a rassemblé certains laboratoires entre eux en fonction de leur expertise et autour d'un thème commun.

Ce chapitre est dédié à la description du projet Two!Ears, description ayant deux buts. Tout d'abord, il s'agit de détailler l'ensemble de l'architecture du logiciel Two!Ears ainsi que son support matériel robotique. Seront ainsi décrits les différents algorithmes et modèles à partir desquels les données sur lesquelles le modèle HTM se base sont issues. D'autre part, cette description expose le cadre dans lequel tout le modèle HTM a été conçu : but du modèle, position dans la chaîne d'analyse des signaux perçus, données accessibles ou non, contraintes temporelles, contraintes conceptuelles, limites techniques etc.

La Sec. 3.1 décrira l'organisation du projet en différents groupes de travail.

La Sec. 3.2 détaillera les principaux composants de l'architecture Two!Ears, notamment le *Blackboard*, le *Scheduler* et les *Knowledge Sources*.

La Sec. 3.3 consistera en une déscription des deux plateformes robotiques utilisées pour la validation expérimentale du logiciel Two!Ears (notre modèle inclus).

La Sec. 3.4 enfin dressera la liste des scénarios de tests créés pour cette validation.

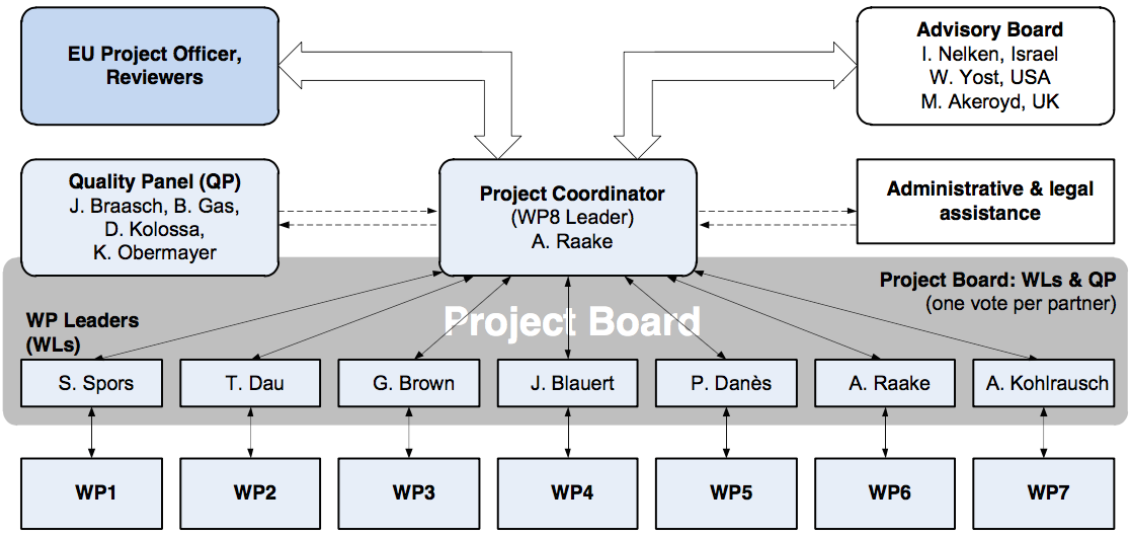


Figure 3.2 – Organisation en Work Packages — Diagramme présentant l'organisation en *Work Packages* du projet Two!Ears, ainsi que les personnes en charge de leur supervision respective (figure d'après [230]).

3.1 Organisation en Work Packages

Comme de nombreux projets européens ou internationaux, le projet est divisé en plusieurs groupes de travail, appelés *Work Packages* (WP). Two!Ears est divisé en sept WP (cf. **Fig. 3.2**), chacun étant dédié à une partie précise de la recherche et de l'implémentation logicielle et/ou matérielle (cf. **Fig. 3.3**) qui sera assignée à un ou plusieurs laboratoires et étant à effectuer selon un calendrier précis. Les sous-sections suivantes décrivent ces sept WP, leurs objectifs ainsi que les tâches respectives qui ont dues être accomplies durant le projet [2].

Work Package 1 (WP1) — Base de données de scénarios. le WP1 est dédié la création et la maintenance d'une base de données audiovisuelles étiquettée, utilisée tout au long du projet pour le développement, l'apprentissage et l'évaluation des modèles perceptifs binauraux implémentés par d'autres WP. Cette base de données contient, en outre :

1. les signaux perçus par les oreilles dont le robot est doté,
2. les réponses impulsionnelles de la tête robotique [3],
3. les enregistrements multicanaux,
4. les réponses impulsionnelles multicanales des pièces dans lesquelles les enregistrements ont été effectués [4],
5. des images fixes,
6. des vidéos.

2. La description des *Work Packages* est principalement tirée de [230].
3. *Head-Related Impulse Responses* — HRIR.
4. *Multichannel Room-Impulse Responses* — MRIR.

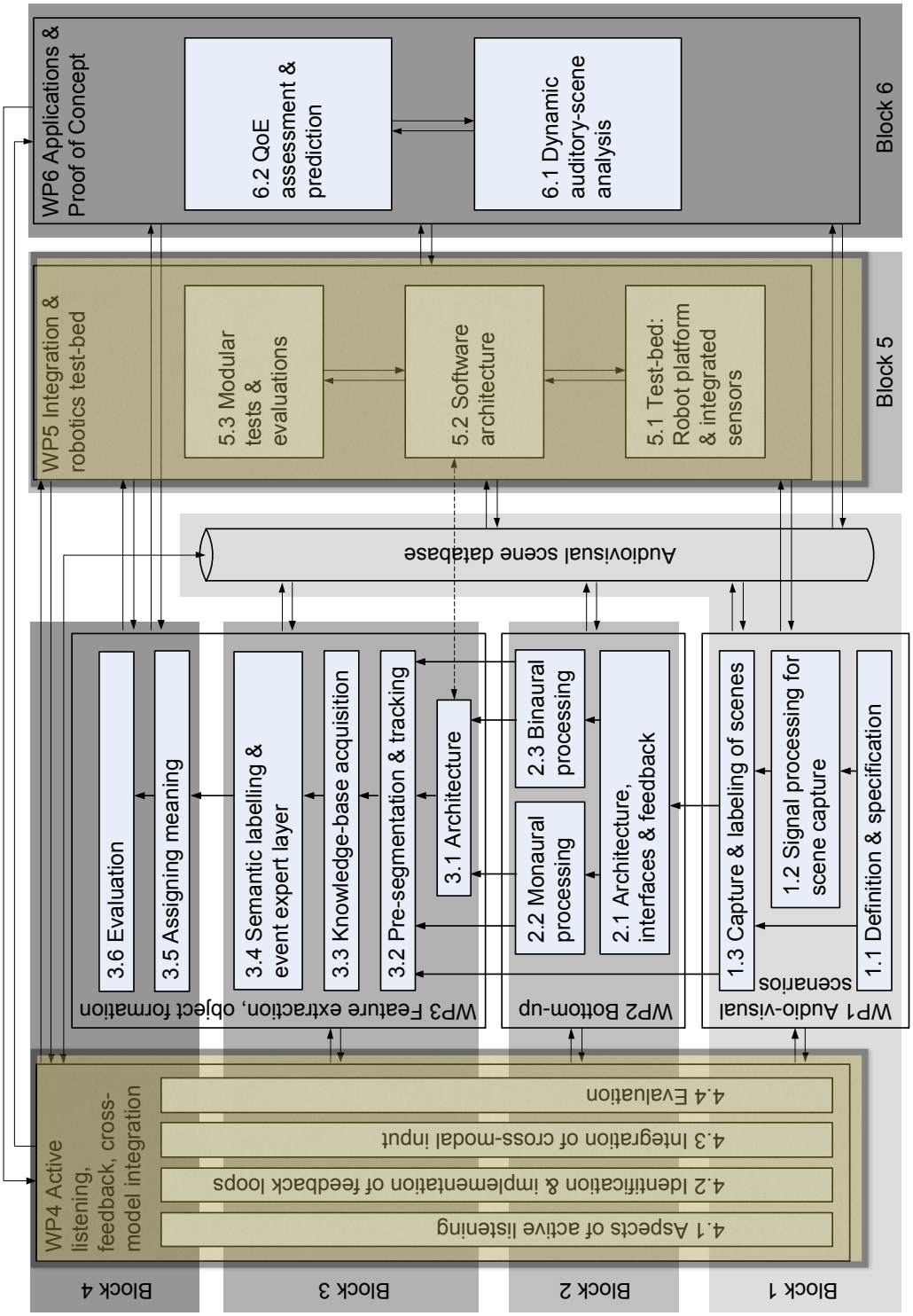


FIGURE 3.3 – RÉSUMÉ SCHÉMATIQUE DU CONTENU DES WP — Connexions entre chaque WP et chaque bloc conceptuel. En jaune pâle sont dénotés les WP dans lesquels ce travail de thèse a été impliqué (figure d'après [230]).

Les principales tâches ont été les suivantes :

1. spécification du format des bases de données et définition des scénarios,
2. analyse de signaux pour la capture et le rendu de scènes acoustiques,
3. collecte, capture, rendu et étiquettage des scènes acoustiques.

Work Package 2 (WP2) — Analyse ascendante des signaux audio. Le WP2 est principalement dédié à l'extraction de caractéristiques des signaux audio perçus par le robot afin de fournir des représentations multi-dimensionnelles de ces signaux, selon des méthodes fortement inspirées du fonctionnement du système auditif subcortical. Le WP2 fournira des versions transformées ou filtrées des signaux audio ainsi que des descripteurs de ces signaux. En fonction de cette analyse, d'autres WP plus haut-niveau pourront requérir une modulation, un raffinement de cette étape d'extraction des caractéristiques bas-niveau des signaux (flux descendant). Les principales tâches ont été les suivantes :

1. définition de l'architecture et de l'interfaçage avec les autres WP,
2. extension de l'étape d'analyse monaurale,
3. extension de l'étape de modélisation binaurale,
4. évaluation.

Work Package 3 (WP3) — Extraction de caractéristiques, formation d'objet et assignation de sens. Le WP3, premier véritable composant de type « expert », a pour but de permettre au système de comprendre et d'analyser des environnements complexes, notamment le cas multi-sources. Trois couches constituent ce WP :

1. la scène auditive est, dans un premier temps, pré-segmentée en différents flux audio correspondant d'un côté aux sources sonores en elles-mêmes (*foreground*) et, de l'autre, aux sons correspondant à l'arrière plan, ou bruit de fond (*background*) ;
2. qualifiée de couche *event-expert*, la deuxième couche va permettre de définir les événements audio qui sont présents dans l'environnement permettant de créer une première description symbolique de la scène, grâce à l'annotation et à l'interprétation sémantique des événements audio perçus ;
3. cette troisième couche permet d'inclure les contraintes contextuelles du scénario dans lequel le système est (scénario S&R par exemple) afin de lever d'éventuelles ambiguités sémantiques.

Les principales tâches ont été les suivantes :

1. définition de l'architecture,
2. pré-segmentation et suivi (*tracking*),
3. aquisition *knowledge-based*,
4. étiquettage sémantique et implémentation de la couche *event-expert*,
5. assignation de sens,
6. évaluation.

Work Package 4 (WP4) — Ecoute active, boucles de rétro-contrôle, intégration d'information inter-modale. Le WP4 est dédié à l'écoute active, processus impliquant de nombreux mécanismes de rétro-contrôle. Ces mécanismes sont le cœur de l'approche descendante (*top-down*) adoptée dans TWO!EARS. Le WP4 se situe au carrefour des autres *Work Packages*, prenant en compte de nombreuses sources d'informations différentes des signaux audio et de leur analyse, comme la position du robot, la direction du torse et de la tête (données proprioceptives et sensorimotrices), ou l'identification d'objets visuels. Les principales tâches ont été les suivantes :

1. recherche bibliographique sur l'écoute active,
2. détermination de boucles de rétro-contrôle impliquant l'audition, la vision et la perception active,
3. implémentation des boucles de rétro-contrôle,
4. intégration de données inter-modales,
5. évaluation.

Work Package 5 (WP5) — Intégration matérielle et logicielle et tests sur le robot. Le WP5 est dédié à l'intégration des modules développés par les WP2, WP3 et WP4 dans le système embarqué du robot. Le but est de pouvoir connecter les modules testés en simulation à une architecture robotique permettant l'utilisation de ces modules avec de vrais signaux et de vrais mouvements. Les principales tâches ont été les suivantes :

1. mise en place de la plateforme robotique,
2. intégration des capteurs audio et visuels,
3. mise en place d'une architecture logicielle permettant à tous les modules computationnels de communiquer avec le robot,
4. tests modulaires et évaluations.

Work Package 6 (WP6) — Applications et preuves de concept. Le WP6 est dédié à l'utilisation concrète du logiciel TWO!EARS. Deux types d'évaluations ont été effectuées :

- analyse dynamique de scène audio : évaluation des performances du logiciel TWO!EARS notamment du point de vue perceptif et cognitif comme, par exemple : reconnaissance de locuteur, reconnaissance de mots-clefs, classification de genre audio etc. Cette évaluation a été faite sur la base des scénarios définis par le WP1 et sur des signaux ayant été préalablement analysés par les WP2-4 ;
- qualité d'expérience : capacité d'experts dédiés à ces tâches de jugement de certaines caractéristiques haut niveau des scènes audio, comme la qualité sonore ou une position d'écoute préférentielle.

Les principales tâches ont été les suivantes :

1. analyse dynamique de scènes audio,
2. qualité d'expérience.

Work Package 7 (WP7) — Dissémination. Le WP7 est dédié à la communication publique autour du projet et à sa diffusion. Cette communication inclut aussi bien une publication intense dans des revues d'excellence et des conférences internationales que la diffusion de l'avancement du projet sur internet *via* la diffusion de vidéos par exemple, sur *Vimeo* ou *YouTube*. Les principales tâches ont été les suivantes :

1. impact scientifique,
2. standardisation des activités,
3. impact industriel.

3.1.1 Discussion

L'ISIR, et ce travail de thèse donc, a fait partie des WP3, WP4 et WP5 mais le travail effectué a été majoritairement dédié au WP4. Le but de celui-ci a été de doter le robot d'une écoute active, c'est-à-dire prenant en compte le résultat d'une analyse d'une scène audio dans la génération de commandes motrices ayant pour but d'améliorer la perception auditive mais également de conférer au robot un comportement réactif. Le travail de ce WP4 a donc consisté en une intégration des actions motrices dans le processus de la perception auditive mais également, de façon spécifique à ce travail de thèse et au modèle HTM, dans une intégration multimodale des données audio et visuelles.

La section suivante va nous permettre d'entrer au cœur du logiciel TWO!EARS par la description de son architecture centrée autour d'un système de type *Blackboard*. Cette architecture a été développée par l'ensemble des membres du consortium, avec des niveaux d'implication différents. Comme la description des WP le montre, le projet TWO!EARS couvre un large spectre de compétences et de domaines d'expertise, résultant en une analyse riche et complexe de l'environnement perçu par un robot. Ainsi, une architecture en même temps robuste et flexible a été indispensable afin de rassembler, organiser et rendre disponible tous les résultats des analyses des différentes entités constituant le logiciel TWO!EARS développé par le consortium.

3.2 Architecture

L'ARCHITECTURE du logiciel TWO!EARS est organisée autour de trois principaux composants : le *Blackboard*, structure dans laquelle sont centralisées toutes les données, que ce soient celles en cours d'acquisition et d'analyse que les données déjà analysées ; les KS, modules « experts » chacun chargés d'une analyse précise des informations reçues par le système ; et le SCHEDULER, responsable de l'exécution des différentes KS composant le système. Les sections suivantes décrivent en détail ces trois principaux composants.

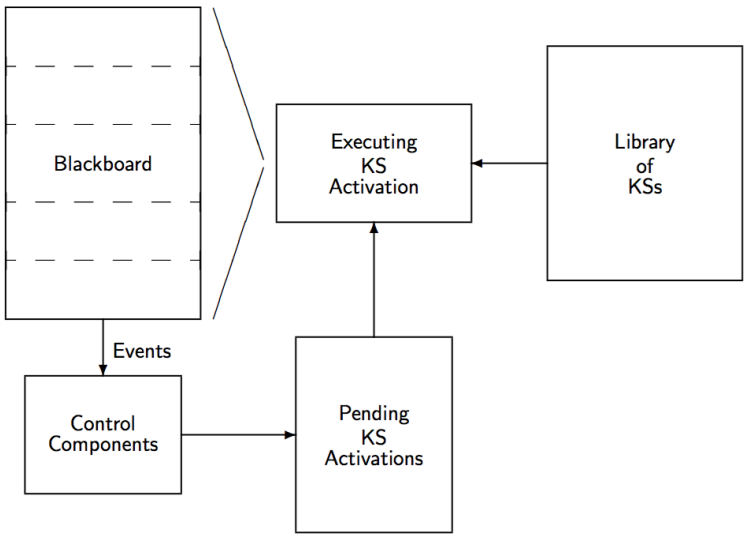


FIGURE 3.4 – SYSTÈME DE TYPE BLACKBOARD — Schéma de la structure générale d'un système de type *Blackboard* (figure d'après [226]).

3.2.1 Système Blackboard

L'organisation du projet TWO!EARS sur la base de l'implémentation d'un système de type *Blackboard* permet, par construction, de fournir une architecture qui intègre la formation d'« expérience » du robot ainsi qu'un comportement actif, à partir de modules individuels fonctionnels. Cette technologie n'est pas neuve : le premier système de type *Blackboard* remonte aux années 80 avec le développement du système *Hearsay-II* par LEE D. ERMAN *et al.* [231]. En 1991, DANIEL D. CORKILL [226] utilise une métaphore pour expliquer le principe d'un tel système :

> « *Imagine a group of human specialists seated next to a large blackboard. The specialists are working cooperatively to solve a problem, using the blackboard as the workplace for developing the solution. Problem solving begins when the problem and initial data are written onto the blackboard. [...] When a specialist finds sufficient information to make a contribution, he records the contribution on the blackboard, hopefully enabling other specialists to apply their expertise. This process of adding contributions to the blackboard continues until the problem has been solved*[5]. »

Un *Blackboard* est donc la mise en commun de plusieurs experts permettant de résoudre de façon coopérative un problème donné en tirant partie de l'expertise de chacune de ces entités et rassemblées autour d'une entité commune. Dans l'implémentation de tels systèmes, un expert est appelé KNOWLEDGE SOURCE (KS).

5. « Imaginez un groupe de spécialistes humains assis à coté d'un grand tableau noir. Les spécialistes travaillent ensemble pour résoudre un problème utilisant ce tableau comme support pour développer leur solution. Cette résolution commence lorsque le problème et les données initiales sont écrites sur le tableau. [...] Lorsqu'un spécialiste a suffisamment d'informations pour proposer une contribution, il l'écrit sur le tableau dans le but de permettre aux autres spécialistes d'appliquer leur expertise. Ce processus d'ajout de contributions au tableau continue jusqu'à ce que le problème soit résolu. »

CORKILL détaille les principales caractéristiques d'un tel système :

Indépendance de l'expertise (« *I think therefore I am* ») : chaque KS doit pouvoir fonctionner sans aide d'une autre KS. Elle puise ses informations du *Blackboard* et résoud une partie du problème en fonction de son expertise.

Diversité des expertises (« *I don't think like you do* ») : le *Blackboard* conçoit chaque KS comme une boîte noire, n'accordant ainsi pas d'importance à la façon dont le problème est traité par la KS : il peut s'agir d'un algorithme utilisant un réseau de neurones, une approche sous forme de règles pré-implémentées etc.

Langage commun (« *What you'd say ?* ») : l'ensemble des connaissances inscrites sur le *Blackboard* doit pouvoir être compris et interprété par tous ses composants. Ainsi, il est nécessaire d'uniformiser la façon dont chaque KS communique les résultats de son expertise.

Métrique positionnelle (« *You could loook it up* ») : dans le cas d'un problème complexe nécessitant des contributions multiples, il devient rapidement nécessaire de filtrer ces contributions en fonction de leur importance relative au problème. En effet, certains résultats peuvent n'être d'intérêt que tard dans le processus de résolution du problème. Ainsi, être capable d'ordonner les contributions des KS permettent d'améliorer l'efficacité computationnelle du système ainsi que sa clarté.

Activation événementielle (« *Is anybody there ?* ») : les KS sont déclenchées par des événements, c'est-à-dire lorsqu'une information nouvelle parvient au *Blackboard*. Et plutôt que de faire scanner le *Blackboard* par les KS de façon incessante, une approche plus efficace consiste à faire indiquer aux KS les événements qui les déclencheront. Ainsi, dès lors qu'un événement survient, le *Blackboard* est capable d'activer la KS concernée.

Contrôle (« *It's my turn* ») : afin d'éviter que plusieurs KS n'interviennent de façon simultanée et désordonnée sur le *Blackboard*, un *manager* est employé afin d'organiser les contributions de chaque KS dans le cas où des conflits surviennent.

Génération incrémentale de la solution (« *Step by step, inch by inch...* ») : la puissance d'un *Blackboard* réside dans leur capacité à résoudre un problème complexe en le traitant de façon incrémentale. Les KS appliquent leur expertise sur un événement venant de se produire et permettent, par publication de leur contribution, d'ajouter un nouvel élément de la solution à trouver. Ainsi, petit à petit, l'ensemble des expertises rassemblées sur le *Blackboard* convergent jusqu'à la solution au problème posé.

Un système de type BLACKBOARD constitue donc un moyen de centraliser le travail d'une communauté d'experts réunis pour résoudre le problème posé initialement. Au sein des sciences informatiques, ce genre d'architecture permet de fournir un support qui homogénéise les contributions des différents experts dédiés à divers types d'analyse computationnelle de données.

La **Fig. 3.4** illustre la structure générale d'un système de type *Blackboard*. Dans le cadre du projet TWO!EARS, le module appelé *Control components* consiste en le *Scheduler*. Ce *Scheduler* est une partie indispensable du *Blackboard* car elle permet d'ordonner et de contrôler les experts afin d'éviter un comportement chaotique du système. La section suivante introduit cette partie essentielle de l'architecture

Blackboard.

3.2.2 Scheduler

Le SCHEDULER est le composant responsable du déclenchement ordonné des différentes KS au sein du *Blackboard*. Cet ordre est déterminé préalablement au lancement du système en fonction des besoins du scénario/contexte dans lequel il va être utilisé, mais sera également dynamiquement recalculé après chaque instance des KS en fonction du résultat de leur analyses. Par exemple, si la tâche que le robot doit effectuer est de se diriger vers toute source sonore correspondant à un signal de parole, dès lors que l'expert d'identification dédié indiquera qu'il a détecté un tel signal, le SCHEDULER pourra redéterminer l'ordre d'exécution des KS favorisant celles permettant d'améliorer la localisation de cette source précisément, ainsi que celles permettant d'effectuer une commande motrice vers la source sonore d'intérêt. Plusieurs facteurs influencent également l'ordre d'exécution des KS :

- chaque KS possède un paramètre appelé *attentional priority*[6]. Les KS avec une priorité élevée sont déclenchées avant celles ayant une priorité plus basse. Cette propriété peut être définie par la KS elle-même ou par des KS externes ;
- toutes les KS possèdent également un paramètre *canExecute* qui permet de vérifier si la KS dispose des éléments nécessaires pour être exécutées :
- les KS sont aussi définies par une fréquence d'invocation maximum qui ne peut être dépassée. Il s'agit d'une fréquence maximum car les KS sont déclenchées par des *événements* et non par un décours temporel. Le SCHEDULER vérifie ainsi que la dernière exécution d'une KS n'a pas eu lieu trop tôt et qu'elle peut ainsi être exécutée à nouveau.

La section suivante décrit en détail les principales KS du système TWO!EARS et qui seront utilisées par le modèle HTM.

3.2.3 Knowledge Sources

Le logiciel TWO!EARS contient une vingtaine de KS. Cependant, cette section ne détaillera que celles qui concernent directement le modèle HTM par souci de concision et de pertinence. L'essentiel de leur formalisation ainsi que la façon dont le modèle interagit avec elles sera également décrit. A nouveau, l'architecture de TWO!EARS permet de récupérer simplement et rapidement toutes ces données, par l'entremise du *Blackboard* dans lequel sont consignées toutes les sorties des KS. A noter que jusqu'à présent nous avons parlé du *modèle HTM* à chaque fois que nous avons décrit les concepts sur lesquels le modèle se base. Nous allons désormais nous référer au modèle en tant que KNOWLEDGE SOURCE, tel qu'il a été *implémenté* au sein du logiciel TWO!EARS. L'implémentation en tant que KS du modèle HTM sera mentionnée en tant : HEADTURNINGMODULATIONKS ou HTMKS. Son architecture ainsi que ses caractéristiques en tant KS seront détaillées plus tard, à la **Sec. 7.2.1**.

6. Priorité attentionnelle.

3.2.3.1 Fonctions auditives

Localisation — DNNLocationKS Une des principales informations portées par les stimuli audio est celle de leur localisation. Tandis que l'humain est capable d'analyser aisément des scènes audio complexes [78], les systèmes computationnels ou robotiques ne sont quant à eux que peu robustes aux conditions acoustiques difficiles, notamment lors de la présence de sources audio interférentes ou d'une réverbération importante. L'algorithme de localisation audio utilisé au sein de TWO!EARS tente d'apporter une amélioration notable des capacités de localisation d'un robot binaural dans des conditions multisources [232]. Pour cela, l'algorithme se base sur des réseaux de neurones profonds [7]. A l'opposé de nombreux systèmes d'audition binauraux, l'algorithme de localisation proposé ici ne se limite pas à une partie de l'espace sonore mais est capable de localiser le son à 360° en azimut. C'est pourquoi, contrairement aux algorithmes traditionnels, l'estimation des ITD et ILD n'est pas suffisant car ces indices binauraux sont identiques pour le champ frontal et le champ postérieur, induisant ainsi des confusions avant-arrière dans les résultats de localisation [233]. Afin de lever certaines incertitudes sur les estimations d'ITD et ILD survenant en conditions réalistes, dues notamment à la réverbération du son provoquant la multiplication d'hypothèses de localisation en azimuth, un apprentissage multi-conditionnel [8] [234, 235] a été effectué.

Les signaux binauraux perçus *via* les oreilles du robot sont d'abord analysés par le WP2 (grâce à l'*Auditory Front-End*, AFE). Cette analyse utilise une banque de 32 filtres Gammatone chevauchant, avec des centres de fréquence répartis uniformément selon une échelle ERB (*Equivalent Rectangular Bandwidth*) s'étendant de 80Hz à 8kHz [237]. L'analyse effectuée par les cellules ciliées internes de l'oreille humaine a été approximée selon la méthode de *half-wave rectification* [9]. Puis la cross-corrélation entre les signaux gauche et droit a été calculée indépendamment pour chacune des 32 bandes fréquentielles sur des trames audio de 20 ms se chevauchant sur 10 ms. De plus, les deux caractéristiques couramment utilisée dans l'analyse de signaux audio, ITD et ILD [78] ont été utilisés. L'ITD est défini comme le décalage correspondant au maximum de la fonction de cross-corrélation ; l'ILD correspond au rapport d'énergie entre les signaux gauche et droit, exprimé en décibel. L'algorithme de localisation proposé ici utilise la fonction de cross-corrélation en entier plutôt que d'estimer l'ITD seulement. Cette approche a été motivée par deux observations :

1. le calcul de l'ITD implique une étape de sélection de pic (maximum de la fonction de cross-corrélation), étape pouvant ne pas être robuste face à des conditions acoustiques diffiles comme lors de la présence de réverbération ;
2. cette fonction de cross-corrélation change systématiquement lorsque l'azimut de la source à localiser change (en particulier, un changement du pic maximum par rapport aux pics voisins).

Pour des signaux binauraux échantillonnés à 16kHz, la fonction de cross-corrélation (CCF) produit un vecteur à 33 composantes pour chaque bande fréquentielle. En y ajoutant la valeur d'ILD calculée également pour chaque bande fréquentielle, le vecteur issu de cette étape d'extraction de caractéristiques est de dimension 34. Ce

7. *Deep Neural Network* — DNN
8. Multi-conditional training
9. Rectification mono-alternance.

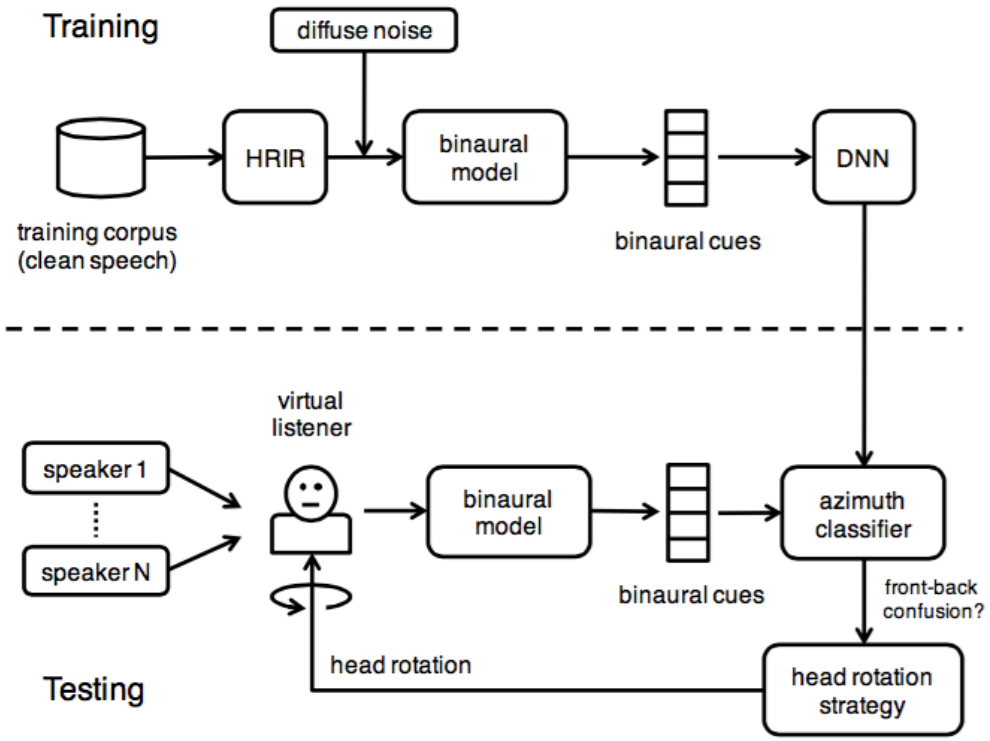


FIGURE 3.5 – SCHÉMA DE L'ALGORITHME D'APPRENTISSAGE DE LOCALISATION AUDIO — (*haut*) Apprentissage, (*bas*) validation. Durant la phase de test, la scène sonore consistant en plusieurs locuteurs est simulée dans un environnement virtuel dans lequel un récepteur binaural est mû afin de simuler la rotation de la tête d'un auditeur humain (figure d'après [236]).

---

vecteur est défini, au temps t et pour une bande fréquentielle centrée autour de la fréquence f, comme $\vec{x}_{t,f}$ A noter qu'à chaque vecteur de caractéristiques a été ajouté du bruit blanc gaussien (variance $\sigma = 0.4$) afin d'éviter le sur-apprentissage et donc de laisser à l'algorithme la capacité de généraliser son apprentissage à des données inconnues. C'est à ce vecteur de caractéristiques $\vec{x}_{t,f}$ que l'algorithme de *DNNLocation* implémenté pour TWO!EARS va tenter de faire correspondre des angles en azimut. Un réseau DNN a été entrainé pour chaque bande fréquentielle [238]. Ces DNN sont des réseaux de neurones à une couche d'entrée, deux couches cachées possédant chacune 128 neurones, et une couche de sortie à $360°/5° = 72$ neurones. La même structure de DNN a été utilisée pour chaque bande fréquentielle afin de ne pas introduire de biais fréquentiel dans l'analyse de l'azimuth. Enfin, la sélection de l'angle « gagnant » est fait selon une décision de type *softmax* [239, 240]. Ces 72 valeurs de sortie des DNN ont été considérées comme des probabilités *a posteriori*, avec $\mathcal{P}(k|\vec{x}_{t,f})$ (où $\sum_k \mathcal{P}(k|\vec{x}_{t,f}) = 1$) où k est l'angle en azimut. Tous les vecteurs de probabilités en sortie de chaque DNN ont ensuite été intégrés en fréquence selon :

$$\mathcal{P}(k|\vec{x}_t) = \frac{P(k)\prod_f \mathcal{P}(k|\vec{x}_{t,f})}{\sum_k P(k)\prod_f \mathcal{P}(\vec{x}_{t,f})} \tag{3.1}$$

où $P(k)$ est la probabilité *a priori* de chaque azimut k. Sachant qu'aucune connais-

sance n'a été introduite dans le système à propos des positions des sources sonores et que toutes les directions ont une probabilité égale d'apparaître, l'**Eq. 3.1** devient :

$$\mathcal{P}(k|\vec{x}_t) = \frac{\prod_f \mathcal{P}(k|\vec{x}_{t,f})}{\sum_k \prod_f \mathcal{P}(\vec{x}_{t,f})} \tag{3.2}$$

D'autre part, la localisation sonore est effectuée pour des morceaux de signaux contenant T trames audio. Ainsi, les probabilités *a posteriori* générées par l'ensemble des DNN a été moyennées sur les T trames selon :

$$\mathcal{P}(k) = \frac{1}{T} \sum_{t}^{t+T-1} \mathcal{P}(k|\vec{x}_t) \tag{3.3}$$

Enfin, la localisation la plus probable est choisie selon la probabilité maximum :

$$\hat{k} = \arg\max_k \mathcal{P}(k) \tag{3.4}$$

Dans des environnements réverbérants et multi-sources, les algorithmes de localisation se trouvent très souvent en difficulté. La réverbération cause une redondance des signaux perçus tandis que la multiplication des sources sonores nécessite une étape préliminaire de ségrégation des flux audio, processus éminemment plus complexe qu'en vision en cela que ces flux sont tous présents, en même temps, dans le signal audio. Afin d'augmenter la robustesse de l'algorithme *DNNLocation*, une phase d'apprentissage multi-conditionnelle (Multi-Conditional Training — MCT) a été effectuée, étape permettant d'augmenter significativement les performances des algorithmes de localisation dans des environnements acoustiques complexes [241, 235, 242]. Les modèles de localisation ont été ici entrainés sur des caractéristiques binaurales MCT créés par mélange d'un signal cible à un azimut spécifique avec un bruit diffus et à différents rapports signal sur bruit (20 dB, 10 dB, 0 dB). Le bruit diffus a consisté en 72 sources de bruit blanc Gaussien, non corrélées, placées dans un plan à 360° par pas de 5°. La base de données utilisée pour l'apprentissage provient de la base TIMIT [243] : 30 phrases sélectionnées aléatoirement pour chacune des 72 positions en azimut auxquelles du bruit diffus a été ajouté (avec un rapport signal sur bruit de 20, 10 et 0 dB).

De la DNNLocationKS, la HeadTurningModulationKS récupèrera non pas la valeur estimée de l'azimut de la source perçue $\hat{k}$ mais le vecteur entier de probabilités *a posteriori* $\mathcal{P}(k)$, après moyennage, selon l'**Eq. 3.3**. Nous désignerons ce vecteur, au temps t, par :

$$\mathbf{\Theta}^a[t] = (\theta_1^a[t], \ldots, \theta_N^a[t])^T \mod 360, \tag{3.5}$$

avec $N = 360/5 = 72$ angles.

Identification — AuditoryIdentificationKS Une autre information importante portée par les signaux, ou les objets audio en fonction du degré d'abstraction

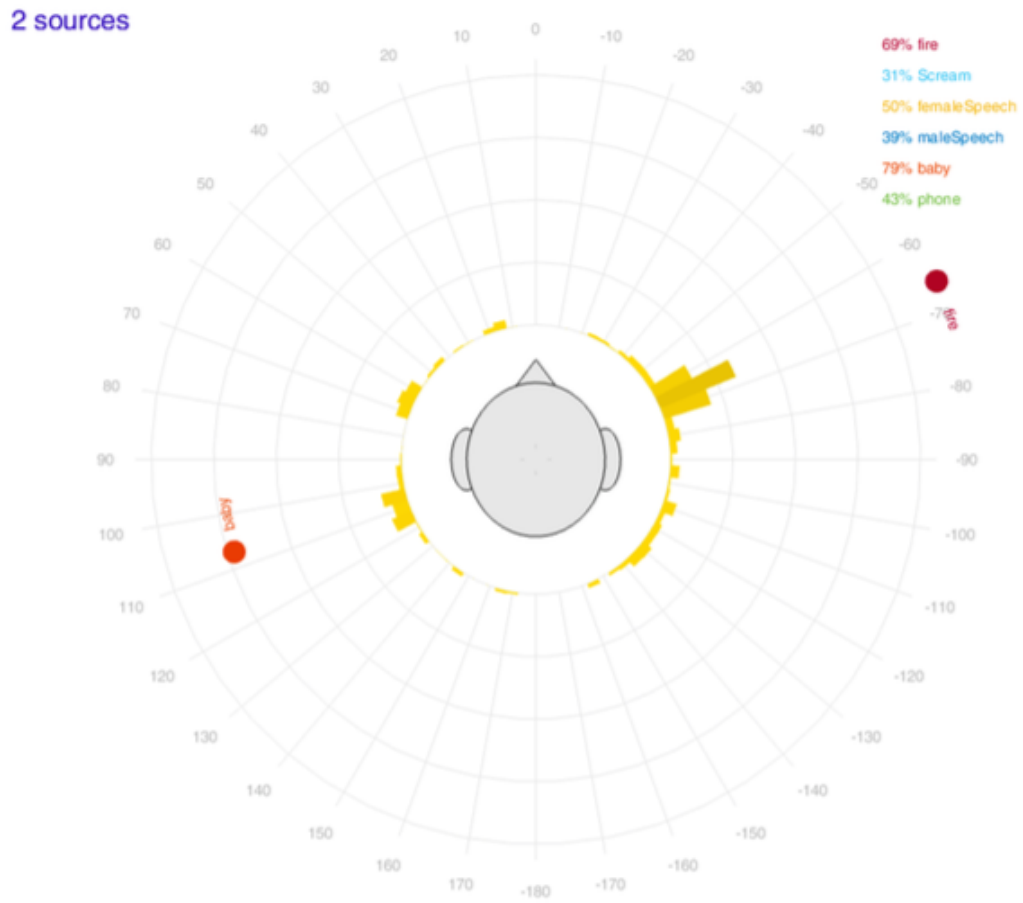


FIGURE 3.6 – IDENTIFICATION ET LOCALISATION DE SOURCES SONORES — Visualisation du résultat de l'analyse du *Blackboard* concernant la localisation et l'identification de deux sources sonores. Les histogrammes jaunes atour de la tête sont les estimations de la localisation des sources ; les scores et les labels associés, en haut à gauche, sont les probabilités résultant de l'analyse par l'algorithme d'identification ; les deux cercles étiquettés correspondent aux objets audio créés à la suite de cette analyse (figure d'après [236]).

---

auquel on se situe, est la *classe* à laquelle ils appartiennent. Cette classe constitue une caractéristique de l'identité de ces sources sonores en cela qu'elle permet de décrire leur contenu d'un point de vue sémantique. Afin de reconnaître l'identité d'une source sonore, des modèles de détection et d'identification ont été créés, capables de déterminer si le flux audio perçu contient un événement sonore particulier ou non. *Détection* et *Identification* sont ici ainsi très proches. Les caractéristiques issus de l'analyse des signaux audio par l'AUDITORYFRONTEND peuvent être de plusieurs types, en fonction du scénario considéré :

- *ratemaps* : Spectrogrammes audio modélisant le taux d'activité du nerf auditif, calculés sur des trames de 20 ms et pour chaque bande fréquentielle Gammatone. Ces *ratemaps* sont créés à partir d'une modélisation des cellules ciliées internes [244, 245, 246, 247] ;
- *caractéristiques spectrales* : 14 différents mesures statistiques ont été utilisées pour « résumer » le contenu spectral de la *ratemap*, à chaque trame de 20 ms, comme la *flatness, kurtosis...* [248, 249, 250, 251] ;
- *onset strengths* : mesuré en décibels à chaque trame temporelle et chaque

bande fréquentielle, calculé à partir de la différence d'énergie entre deux *ratemaps* [252] ;

- *spectrogramme de modulation en amplitude* : chaque bande fréquentielle correspondant à la modélisation des cellules ciliées internes est analysée selon une banque de filtres de modulation selon une échelle logarithmique [253, 254]. A chaque trame temporelle a été assignée (x bandes fréquentielles) $\times$ (y valeurs de modulation des filtres) ;
- *caractéristiques de Gabor* : détecteurs de bords sensibles à l'orientation spectro-temporelle à partir de la représentation en *ratemap* [255].

Chaque AUDITORYIDENTITYKS est dédiée à une classe audio, par exemple : {`speech`}, {`knock`}, {`alarm`}. La HEADTURNINGMODULATIONKS récupère les probabilités d'appartenance à chaque classe audio disponible et les concatène en un seul vecteur, sans les normaliser (les probabilités étant indépendantes entre elles). Ce vecteur sera noté, au temps discret t, comme suit :

$$\mathbf{P}^a[t] = (p_1^a[t], \ldots, p_{N_a}^a[t])^T \tag{3.6}$$

avec $p_i^a[t]$ étant la probabilité que la trame audio t appartiennent à la i-ème catégorie audio. La **Fig. 3.7** illustre les résultats de l'identification des classes de sons que nous utiliserons lors de l'évaluation du modèle HTM sur la plateforme robotique de l'ISIR. Ces résultats ont été mesurés en condition réelles, dans la salle d'expérimentation du bâtiment ADREAM du LAAS, à Toulouse. Nous voyons notamment que la plupart des classes audio sont bien reconnues (selon le critère de *balanced accuracy*, BAC [256]), à l'exception des sons de type *crash*. Cependant, nous verrons lors de l'évaluation du modèle HTM en conditions réelles sur notre plateforme robotique que les résultats de classification des experts d'identification audio sont globalement nettement moins bons (aux alentours de 40%). Cela ne gênera pas notre modèle, bien au contraire puisqu'il a l'ambition de pouvoir gérer correctement ces erreurs de classification.

3.2.3.2 Fonctions visuelles

Contrairement aux experts dédiés à l'analyse du signal audio, le système visuel dont le robot est doté envoie toutes les informations dans une même structure de données. La HEADTURNINGMODULATIONKS va ainsi récupérer séparément les données visuelles de localisation et d'identification mais les traiter de façon similaire aux KS dédiées au signal audio. Les KS présentées ici ont été implémentées par l'ISIR selon la même structure que les autres KS, dans un but de cohérence et d'intégration optimale.

Identification Un algorithme de détection et de traque de personnes a été implémenté dans le cadre de TWO!EARS. Cependant, il n'a été que très peu utilisé dans le cadre du projet. Ainsi, seul l'algorithme d'identification des objets sera décrit dans cette section. L'identification visuelle se base sur l'algorithme *Linemod* [258, 259], présent dans les bibliothèques *OpenCV*[10]. Il s'agit d'un algorithme de détection

10. `http://opencv.org/`

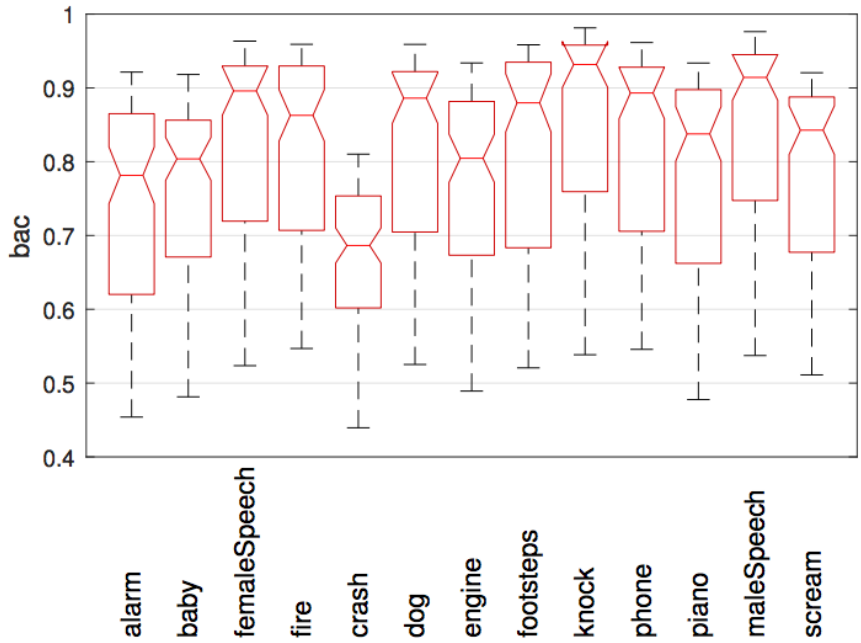


FIGURE 3.7 – PERFORMANCE DANS L'IDENTIFICATION AUDIO — Résultats de l'identification des sons que nous utiliserons lors de l'évaluation du modèle sur la plateforme robotique de l'ISIR. Ces résulats ont été obtenus dans le bâtiment ADREAM du LAAS, à Toulouse. Le critère BAC, pour *balanced accuracy*, a été utilisé en tant que mesure de performance et consiste en une moyenne arithmétique de la *sensibilité* (vrais positifs) et de la *spécificité* (vrais négatifs) de l'algorithme de classification (figure d'après [257]).

---

d'objet multimodal [11] utilisant principalement des gradients de couleur et de surface calculés à partir d'un nuage de points RGB-D [12] (cf. **Fig. 3.8**).

L'algorithme *Linemod* a été développé pour détecter des objets sans *texture* dans des environnements encombrés. L'approche employée est de type *template-matching* : une étape d'apprentissage initiale consiste en l'acquisition des points en RGB-D à partir de plusieurs points de vue (distances et angles divers) et en la détection des points communs existant afin de créer un modèle (*template*) de l'objet à reconnaître. Ainsi, le système dispose d'une base de données de modèles d'objets qui seront utilisés par la suite pour la détection et l'identification d'objets visuels par recherche du modèle le plus ressemblant. Une fois cet apprentissage hors-ligne effectué, un paquet ROS dédié s'occupe de la détection en temps réel des objets modélisés dans les images acquises par les caméras. Ainsi, à chaque trame t et pour chaque objet modélisé — et contrairement aux algorithmes d'identification audio utilisé dans TWO!EARS — une valeur binaire $P^v[t] \rightarrow \mathbb{N} \in [0, 1]$ est attribuée selon la détection ou non de l'objet dans l'image en question.

Afin de conserver une unité dans la façon dont les informations sont collectées par le modèle HTM, la VISUALIDENTITYKS implémentée pour gérer les données issues de la reconnaissance d'objets visuels rassemble, à chaque trame t, chacune des valeurs

11. La notion de multimodalité est ici à différencier de la multimodalité sensorielle (audiovisuelle par exemple) telle que nous l'avons comprise jusqu'à présent. Ici, une modalité signifie une source d'information distincte.

12. *Red Green Blue - Depth* (rouge vert bleu - profondeur).

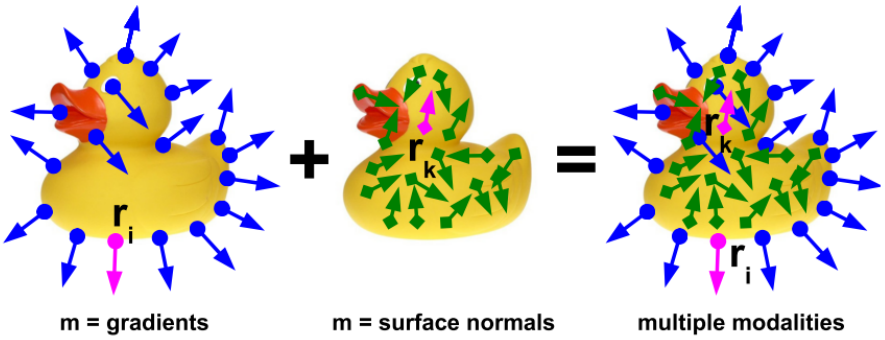


FIGURE 3.8 – ALGORITHME DE RECONNAISSANCE VISUELLE — Illustration de l'algorithme *Linemod* effectuant l'identification d'objets visuels. Un objet est défini selon plusieurs modalités. (*gauche*) les gradients de l'image sont principalement déterminés à partir des contours de l'objet, (*milieu*) normales de surface déterminés par le corps de l'objet, (*droite*) combinaison des deux. L'approche *Linemod* combine ces deux informations afin d'émettre une hypothèse sur la présence d'un objet (figure d'après [258]).

$P^v[t]$ dans un seul vecteur, selon :

$$\mathbf{P}^v[t] = (p_1^v[t], \ldots, p_{N_v}^v[t])^T, \tag{3.7}$$

Ainsi, de la même manière que pour les AUDITORYIDENTITYKS, chaque objet modélisé est considéré comme un expert dédié à la reconnaissance d'un objet en particulier. Chacun de ces experts d'identification visuelle émet donc une probabilité d'appartenance à une classe visuelle. Ainsi, et comme dans l'**Eq. 3.6**, p_i^v représente la probabilité que la n-ième trame visuelle appartienne à la i-ème catégorie visuelle.

Contrairement aux experts d'identification audio, ceux dédiés à la reconnaissance visuelle ne font que peu voire pas d'erreurs du tout. Malgré cela, lors de l'élaboration du simulateur destiné à évaluer le modèle HTM, nous avons inclus un taux d'erreur de ces experts égal à celui des experts audio afin de garantir que le modèle ne se base pas préférentiellement sur une modalité plutôt qu'une autre pour gérer les erreurs de classification.

Localisation La position des objets détectés est incluse dans l'algorithme de reconnaissance/détection des objets (cf. **Sec. 3.2.3.2**). Contrairement à l'identification audio, l'algorithme de reconnaissance visuelle ne renvoie pas un vecteur de probabilité mais une seule valeur de localisation par objet détecté. La VISUALIDENTITYKS implémentée pour récupérer les données de localisation visuelle selon un mode similaire aux autres KS, communique également avec le *Blackboard* afin de récupérer la position du torse (ou de la base, ces deux positions étant les mêmes) et celle de la tête. La position du torse sera utilisée afin d'exprimer les angles de détection des objets dans un référentiel absolu, de la même façon que les angles audio (dans le cadre du modèle HTM) Les données de la VISUALLOCATIONKS se présentent donc sous la forme :

$$\Theta^v[t] = \theta^v[t] + \theta_{head} + \theta_{torso} \mod 360 \tag{3.8}$$

3.2.4 Regroupement

Les sorties de ces quatre KS seront finalement collectées dans un vecteur unique $\mathbf{V}[t]$ défini comme :

$$\mathbf{V}[t] = (\mathbf{P}[t]^T, \mathbf{\Theta}[t])^T \tag{3.9}$$

avec

$$\mathbf{P}[t] = (\mathbf{P}^a[t]^T, \mathbf{P}^v[t]^T)^T, \tag{3.10}$$
$$\mathbf{\Theta}[t] = (\mathbf{\Theta}^a[t]^T, \mathbf{\Theta}^v[t])^T \tag{3.11}$$

A noter que pour la localisation visuelle, nous avons intégré l'angle $\Theta^v[t]$ dans un vecteur de localisation visuelle similaire à celui généré par l'expert de localisation audio. Ainsi, reprenant la valeur de $N = 360/5 = 72$ angles de localisation par pas de 5°, le vecteur de localisation que nous utiliserons sera exprimé selon :

$$\mathbf{\Theta}^v[t] = (\theta_1^v[t], \ldots, \theta_N^v[t]) \tag{3.12}$$

avec

$$\theta_i^v[t] = \begin{cases} \theta_i^v[t] \text{ si } i = N_i, \\ 0 \text{ sinon} \end{cases} \tag{3.13}$$

où N_i est la valeur de l'angle de localisation visuelle. Dans tout ce qui suit, la « n-ième trame audiovisuelle » fera référence au vecteur $\mathbf{P}[t]$ tandis que la « n-ième localisation audiovisuelle » fera référence au vecteur $\mathbf{\Theta}[t]$. Le vecteur $\mathbf{V}[t]$ sera ensuite récupéré par le HeadTurningModulationKS à partir duquel toutes les analyses pourront être effectuées. A noter que le modèle HTM se base uniquement sur les sorties des experts de classification et de localisation. En aucun cas il n'analyse les caractéristiques bas-niveau des signaux tout comme il ne peut les modifier.

3.2.5 Discussion

Cette section a décrit en détail l'architecture du système Two!Ears et ses trois principaux composants : le Blackboard, le Scheduler et les Knowledge Sources. Les systèmes de type *Blackboard* sont couramment utilisés en robotique lorsqu'il s'agit de mettre en commun les résultats de nombreuses analyses d'une partie fragmentée d'un problème global à résoudre. Ici, le problème général à résoudre est la compréhension d'un environnement audiovisuel et l'utilisation de la représentation qui en découle à des fins de navigation modulée par une tâche précise à accomplir (scénario S&R par exemple). Ce problème global nécessite l'analyse préalable de ses nombreuses parties : traitement des signaux audio et visuels, identification & localisation audio et visuelle, cartographie SLAM puis navigation, détermination

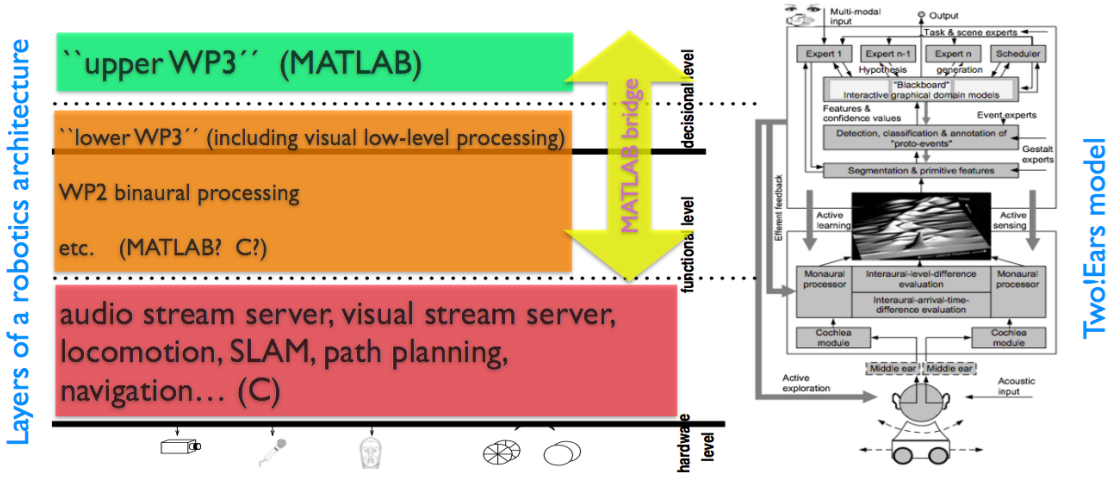


FIGURE 3.9 – De la modélisation computationnelle du projet TWO!EARS (*droite*) à une architecture logicielle robotique temps réel (*gauche*) (figure d'après [260]).

de la tâche à effectuer, prise de décision en fonction de cette tâche etc. Chaque KS dédiée à la résolution de ces sous-problèmes peut ainsi profiter du support de type BLACKBOARD pour récupérer les données dont elles ont besoin puis communiquer les résultats issues de leurs analyses. La HEADTURNINGMODULATIONKS, en tant qu'implémentation du modèle HTM selon les règles dictant le développement des KS, est inclus dans ce *framework* et est soumise aux mêmes contraintes.

D'autre part, nous avons passé en revue les KS dont la HEADTURNINGMODULATIONKS a besoin : DNNLOCATIONKS, AUDITORYIDENTITYKS, VISUALLOCATIONKS et VISUALIDENTITYKS, principalement.

La section suivante introduit les deux robots sur lesquels le système TWO!EARS a été porté mais décrira majoritairement le robot du LAAS : Jido. En effet la majeure partie du travail d'intégration s'est faite sur ce robot. Cependant, une étape significative d'adaptation, de tests et de validation du fonctionnement du système TWO!EARS sur Odi, le robot de l'ISIR, a été indispensable et a fait partie de ce travail de thèse. Ce travail sera décrit plus tard, à la **Sec. 7.2**.

3.3 Les robots — Jido & Odi

L'ENSEMBLE du logiciel TWO!EARS a pour but d'être intégré à la plateforme robotique mobile. Deux plateformes robotiques ont été à disposition du projet : l'une au LAAS, à Toulouse, l'autre à l'ISIR, à Paris. Le travail effectué sur ces deux robots a essentiellement fait partie du WP5. Les buts du WP5 ont été multiples :

1. fournir une plateforme matérielle mobile et dotée de capteurs audio et visuels ;
2. fournir la plateforme logicielle permettant de récupérer les données issues du robot, que ce soit l'odométrie ou les signaux perçus par les capteurs ;

FIGURE 3.10 – TÊTE ET TORSE KEMAR — Les robots Jido & Odi ont été dotés d'un *Head and Torso Simulator* (HATS) reproduisant un certain type de morphologie humaine. Cette tête dispose de deux oreilles au sein desquelles se trouvent deux micros. De plus, le cou a été augmenté avec un dispositif permettant des rotations de la tête. Enfin, un dispositif de vision binoculaire a également été ajouté pour Jido (robot du LAAS), tandis qu'une seule caméra a été ajoutée à Odi (robot de l'ISIR). Ce mannequin a été fixé sur une base mobile permettant la navigation et la cartographie de l'environnement.

3. intégrer l'ensemble des développements des WP2-4, au sein du robot et selon une architecture simple et aisément utilisable.

Cela a impliqué le développement de trois plateformes de test :

a. un simulateur [13] de tête et de torse anthropomorphique et binaural [14] ayant un degré de liberté en azimut au niveau du cou (cf. **Fig. 3.10**) ;
b. le même système mais auquel une vision stéréoscopique a été ajoutée ;
c. le montage d'une tête binaurale sur le robot mobile.

De plus, une architecture logicielle modulaire a été fournie avec ce support matériel. Cette architecture est composée d'une couche fonctionnelle bas-niveau faite de composants exécutés sous de fortes contraintes temporelles et communiquant en temps réel. Le développement d'un pont entre le langage MATLAB®, intensivement utilisé par l'ensemble des WP, et cette architecture logicielle développée en langage C a également été fournie par le WP5. Grâce à lui, il est possible de communiquer directement *via* MATLAB® avec le robot ainsi que de gérer les différentes contraintes temporelles entre les couches « cognitives », à la dynamique lente, et les couches « robotiques », à la dynamique temporelle beaucoup plus élevée.

Dans un premier temps, la **Sec. 3.3.1** exposera la plateforme robotique d'un point de vue matériel. Ensuite, la **Sec. 3.3.2** décrira les composants logiciels liés au robot.

3.3.1 Description matérielle

3.3.1.1 Partie mobile

La plateforme robotique de Toulouse, Jido, possède une base mobile MP-700 de Neobotix [15]. Il s'agit d'une plateforme mobile dotée d'une paire de roue fonctionnement

13. Le terme « simulateur » est ici synonyme de « mannequin ».
14. HATS : *Head-And-Torso Simulator*.
15. `http://www.neobotix-robots.com/mobile-robot-mp-700.html`

selon un mode différentiel non-holonomique (les roues sont situées de chaque côté du robot et sont indépendantes). Une troisième roue, non motorisée, est située à l'arrière du robot afin d'améliorer sa stabilité horizontale. La charge maximum que la base mobile peut supporter est de 300kg, permettant ainsi de monter la tête et le buste KEMAR. Les moteurs ainsi que leur encodeurs relatifs sont connectés à un contrôleur *Harmonica*[16] similaire à ceux utilisés pour le moteur du cou du HATS. De plus, Jido embarque deux LASER de type SICK LMS200[17], un à l'avant, un à l'arrière. Enfin, le robot est équipé d'un ordinateur sans ventilateur (afin d'éviter la contamination des signaux audio) doté d'un processeur Intel® CoreTM i7 CPU E610 cadencé à 2.53Ghz et possédant 4GB de RAM. Deux interfaces BUS CAN ont été également ajoutées afin de connecter les capteurs et les actuateurs. Sur la base mobile est monté un *Head-And-Torso Simulator* (HATS) anthropomorphique : le modèle *KEMAR Type 45BB-2* (cf. **Fig. 3.10**). Ce modèle de tête et torse est également doté de deux larges oreilles[18], dans lesquelles sont placées, à l'intérieur de la tête, deux microphones *G.R.A.S. Type 26CS*[19]. La tête du robot est également dotée de mouvements de rotation horizontale grâce à l'ajout d'un cou, d'un moteur, d'un encodeur et d'un micro-contrôleur (cf. **Fig. 3.11**). Une partie logicielle en charge du contrôle de la position et de la vélocité du cou a été encapsulé dans un composant *GenoM3* dédié.

La vélocité angulaire de la tête, par rapport au torse et à la base mobile, est notée ω_{head}. D'autre part, pour certaines applications comme la localisation active, il est nécessaire que le vecteur de vélocité de la tête *KEMAR*, incluant sa position en (x, y, z) soit exprimé dans son propre référentiel : (H, x_H, y_h, z_h) (cf. **Fig. 3.11**). Ainsi, sont définies les notations suivantes :

v_y : la vélocité linéaire de la tête le long l'axe inter-aural ;
v_z : la vélocité linéaire de la tête par rapport à la direction de mise au point (*boresight direction*) ;
ω_x : la vélocité angulaire autour de l'axe vertical.

Ces vélocités sont relatives au monde mais sont exprimées dans le référentiel de la tête. Un contrôle omnidirectionnel de la tête (v_y, v_z, ω_x) implique un mouvement conjoint de la base et du cou, par l'entremise des trois paramètres de contrôle $(v_{base}, \omega_{base}, \omega_{head})$. Cela conduit à ce que la base « suive » la rotation de la tête lorsque c'est nécessaire.

L'**Eq. 3.14** permet de déterminer les commandes théoriques de vélocité $(v_{base}, \omega_{base}, \omega_{head})$ à utiliser afin d'atteindre n'importe quel vecteur de vélocité (v_y, v_z, ω_x) :

$$\begin{bmatrix} v_y \\ v_z \\ \omega_x \end{bmatrix} = \underbrace{\begin{bmatrix} -\sin q & D\cos q & 0 \\ \cos q & D\sin q & 0 \\ 0 & -1 & -1 \end{bmatrix}}_{J(q)} \begin{bmatrix} v_{base} \\ \omega_{base} \\ \omega_{head} \end{bmatrix} \tag{3.14}$$

où q est l'angle de la tête par rapport au torse et D est la distance entre le milieu

16. http://www.elmomc.com/products/harmonica-main.htm
17. http://sicktoolbox.sourceforge.net/docs/sick-lms-technical-description.pdf
18. http://www.gras.dk/45bb-2.htmlandhttp://www.gras.dk/media/docs/files/items/m/a/man_45BB_45BC.pdf
19. http://www.gras.dk/26cs.html

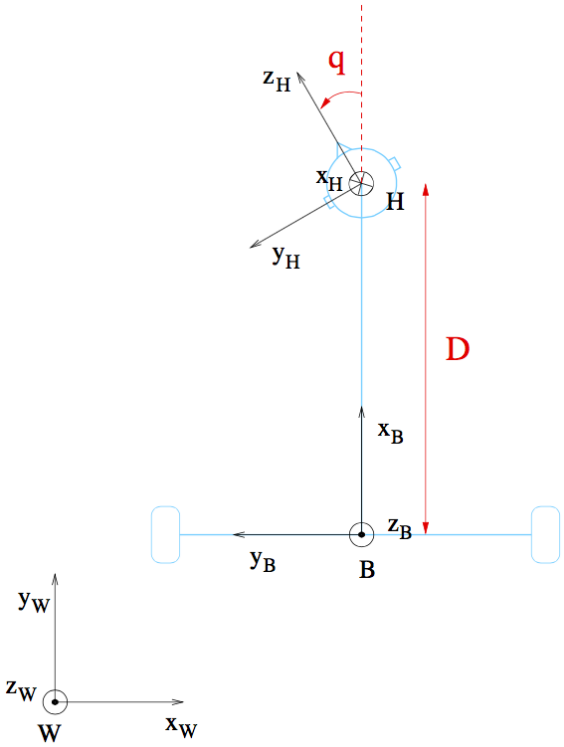


FIGURE 3.11 – VUE DU ROBOT, DE HAUT — (*bleu*) la base et la tête ; (*W*) *World* ; (*B*) Base mobile ; (*H*) *Head*. Les points B et H sont définis ici dans un plan horizontal commun avec *D* comme la distance entre ces deux points.

---

de l'axe inter-roues et l'axe de rotation de la tête. Cependant, certaines limites empêchent la rotation complète de la tête, en particulier la contrainte de rotation similaire à celle d'un cou humain. Ainsi, il a été rendu impossible, *via* des limites logicielles, de tourner la tête à gauche ou à droite à plus d'un certain angle.

3.3.1.2 Partie auditive

La gestion des signaux audio binauraux aquis par les deux microphones (cf. **Sec. 3.3.1**) est assurée par le composant *GenoM3 Binaural Audio Stream Server*[20] (*BASS*). BASS permet de transmettre des données audio provenant de n'importe quel matériel compatible avec ALSA (*Advanced Linux Sound Architecture*[21]) vers d'autres composants de l'architecture logicielle. BASS permet de paramétrer, démarrer et stopper l'acquisition ainsi que d'envoyer les données capturées vers un port de sortie. Dans son mode *capture*, le matériel audio envoie périodiquement un nombre de données au composant BASS. Ce nombre, généralement exprimé en nombre de trames, est fixé avant l'aquisition des données. BASS envoie ensuite ces données vers son port de sortie, agissant comme une mémoire tampon : il est par exemple possible de le configurer en mode FIFO[22] contenant les deux dernières secondes de signal

20. Disponible à l'adresse `https://github.com/TWOEARS/audio-stream-server`
21. ALSA fait partie du noyau de Linux, fournissant ainsi des *drivers* pour les matériels audio. Voir `http://www.alsa-project.org`
22. First In First Out — Premier entré, premier sorti.

FIGURE 3.12 – Caméras μeye UI-3241LE-C-HQ. (*gauche*) vue avant ; (*droite*) vue arrière.

aquis.

3.3.1.3 Partie Visuelle

Cette section est consacrée à la description de la partie matérielle et logicielle du système visuel du robot.

Caméras. Les caméras choisies pour le système de vision stéréoscopique sont des μeye UI-3241LE-C-HQ [23] de la marque IDS, se connectant *via* une interface USB 3.0 (cf. **Fig. 3.12**). Les caméras sont dotées de capteurs CMOS de $1,3$ megapixels, 1/1.8“, à une résolution de 1280×1024. Deux options de capture sont disponibles, ayant un effet sur l'ouverture du diaphragme : *rolling shutter* pour les environnements très peu bruités et au contraste élevé, ou *global shutter* si des objets sont en mouvement. La taille de la lentille est de $6,784$ mm $\times$ $5,427$ mm et la taille d'un pixel est de $5,3$ µm. La taille de chaque μeye est de $36,0$ mm $\times$ $36,3$ mm $\times$ $20,2$ mm [24] et leur poids est de 12 g. Une API [25] est fournie par le constructeur ainsi que les *drivers* nécessaires. Y sont incluses toutes les fonctions permettant d'avoir accès aux paramètres des caméras ainsi que celles permettant leur configuration.

ROS. Afin d'intégrer la partie logicielle dédiée aux caméras μeye à l'architecture TWO!EARS, le paquet libre ROS *ueye* [26] a été utilisé. Ce paquet fournit un nœud *driver* pour avoir accès aux images perçues par les caméras, images qui seront publiées sur un topique ROS dédié.

Calibration. Afin de calibrer la paire de caméra montée en configuration stéréoscopique, le paquet *stereo_image_proc* [27] a été utilisé. Ce paquet utilise les *drivers* des caméras afin d'acquérir des images et d'envoyer le flux aux nœuds ROS chargés de leur analyse. Il permet également d'effectuer une rectification des images stéréoscopiques brutes consistant en la projection des images dans un plan commun afin que

23. `https://en.ids-imaging.com/store/ui-3241le.html`
24. Hauteur x Largeur x Longueur.
25. `https://en.ids-imaging.com/manuals/uEye_SDK/EN/uEye_Manual/index.html?c_programmierung.html`
26. `http://wiki.ros.org/ueye`
27. `http://wiki.ros.org/stereo_image_proc`

FIGURE 3.13 – CALIBRATION DES CAMÉRAS — Echiquier géant utilisé lors de la calibration des caméras en configuration stéréoscopique.

les coordonnées de chaque image issue des caméras soient les mêmes. La calibration des caméras est sur un algorithme classique de calibration [28] utilisant un modèle dit de *pinhole camera* disponible sous *OpenCV*[29]. Selon ce modèle, une vue de la scène est formée par projection des points en 3D sur une image plane par transformation de perspective. Les paramètres intrinsèques et extrinsèques des caméras sont obtenues en bougeant un objet à la géométrie connue et dont les points caractéristiques sont facilement détectables (cf. **Fig. 3.13**). A noter qu'*OpenCV* possède un mode dédié pour la calibration basée sur un objet référence tel que l'échiquier géant utilisé ici.

3.3.2 Description logicielle

3.3.2.1 Robot Operating System — ROS

ROS est une plateforme logicielle largement répandue dans la communauté robotique. Il s'agit non seulement d'un *middleware* mais également d'une large bibliothèque de fonctionnalités implémentées en tant que composants logiciels (comme la localisation, la cartographie, la planification de chemins, l'évitement d'obstacles etc.), et ce avec une utilisation et une installation faciles et rapides. ROS dispose d'une architecture se basant sur le principe de composants logiciels permettant (i) leur exécution de façon concurrente et distribuée, (ii) la réutilisation du logiciel et (iii) la rapidité des tests. La principale terminologie de ROS est résumée ici :

Nœuds : les nœuds sont les composants logiciels utilisant ROS.

Messages et topiques : les flux de données sont appelés topiques. Un nœud qui publie des données le fait dans un topique ; un nœud qui nécessite des données

28. `http://docs.opencv.org/modules/calib3d/doc/camera_calibration_and_3d_reconstruction.html`
29. `http://opencv.org/`

souscrit à un topique. Les données transitant *via* les topiques sont appelés des messages. Chaque message est constitué de plusieurs champs de données formant ainsi une structure de données appelée *message type*. Sachant qu'un topique ne porte qu'un seul type de message, les termes *topic type* et *message type* sont utilisés indifféremment.

Services et actions : nœuds pouvant fournir des *services* afin de les contrôler. Un service peut prendre des paramètres en argument lorsqu'il est invoqué et peut également retourner des paramètres une fois l'exécution terminée. Les services dont la durée d'exécution est longue (comme l'acquisition d'images) sont plutôt définis comme des *actions* incluant des mécanismes de *feedback* durant leur exécution.

Un logiciel utilisant ROS est organisé en *paquets*. Un paquet contient des nœuds ROS, mais aussi des bases de données, des fichiers de configuration etc. Les paquets ROS sont eux-mêmes organisés en *stacks*[30] permettant d'utiliser ROS de façon distribuée. Par exemple, le *navigation stack* contient de nombreux paquets dédiés à la navigation d'une base mobile dans une carte de l'environnement apprise au préalable. Le système d'exploitation officiellement supporté à ce jour par ROS est *GNU/Linux Ubuntu*. Différentes version de ROS existent avec des restrictions de compatibilité concernant les version d'*Ubuntu*. Le robot fonctionne sous la dernière version à support étendu d'*Ubuntu 14.04*.

3.3.2.2 GenoM

GenoM (pour *Generator of Modules*) est un outil permettant de construire des architectures logicielles fontionnant temps réel, notamment pour des sytèmes complexes embarqués (satellites, robots autonomes etc). Développé par le LAAS, GenoM[31] est un générateur de composants modulaires indépendants permettant notamment de faire communiquer une entité logicielle externe avec les logiciels internes au robot. GenoM3, la version actuelle, permet, entre autres :

- l'intégration de fonctions hétérogènes possédant des contraintes temps-réel et des complexités algorithmiques différentes (contrôle des capteurs et des actuateurs, analyse des données, planification des tâches etc.) ;
- l'intégration homogène de ces fonctions dans une architecture de contrôle qui requiert un comportement cohérent et prédictif (démarrage, terminaision et gestion d'erreurs par exemple) et une interface standardisée (configuration, flux de contrôle, flux de données) ;
- la gestion de la parallélisation, de la distribution physique ainsi que de la portabilité de ces fonctions ;
- une simplicité d'utilisation pour les non-spécialistes (ajout, modification ou (ré)utilisation des fonctions).

GenoM3 génère ainsi le code source des composants demandés en utilisant un gabarit générique commun à tous les composants. Ce gabarit est externe à GenoM : des gabarits différents peuvent ainsi être développés. Enfin, à l'utilisation d'un gabarit

30. En « tas ».
31. `https://git.openrobots.org/projects/genom3`

unique pour tous les composants est couplée une description formelle de l'interface des composants basée sur le langage OMG IDL. Le projet GenoM est libre de droit et sous la licence BSD-like.

3.3.2.3 Matlab-to-Ros bridge

L'interfaçage de MATLAB® et ROS a été un besoin exprimé tôt durant le projet. En effet, toutes les KS étant développée en MATLAB® et le robot fonctionnant sous ROS, il a fallu trouver un moyen de faire communiquer les deux architectures de façon rapide, simple et robuste. Cette interface n'est pas un besoin exclusif au projet TWO!EARS : plusieurs solutions ont déjà été développées par la communauté robotique mais aucune ne s'est réellement imposée en raison souvent d'une complexité d'utilisation ou d'installation. Trois approches principales ont émergé pour résoudre ce problème :

l'approche MEX permettant d'inclure l'API ROS en C++ au sein de fichiers MEX. Cette approche aboutit souvent à des erreurs de compilation et d'exécution dues à des incompatibilités entre les versions des nombreuses bibliothèques utilisées (comme *boost* par exemple) et les compilateurs C++ utilisés par ROS et MATLAB®. Cela pourrait être résolu en reconstruisant tout ROS avec les mêmes versions de compilateurs et de bibliothèques que celles utilisées par MATLAB®.

l'approche Java se basant sur l'API ROS Java, permettant d'écrire du code en langage Java dans MATLAB®. Bien que non supportée officiellement par ROS, cette approche est performante et suffisamment mâture pour être utilisée avec sécurité. De plus, elle permet une bonne intégration entre MATLAB® et ROS, grâce la conversion automatique des types de données de Java dans l'espace de travail MATLAB®. Cette solution est cross-plateforme.

l'approche « pont » utilisant une interface logicielle entre ROS et MATLAB®. Ici, ROS et MATLAB® ne sont pas directement connectés. Cependant, cette approche permet une plus grande adaptabilité et peut facilement être cross-plateforme. La communication avec MATLAB® se fait *via* un protocole de type TCP-IP.

Bien que MATHWORKS™ ait sorti une solution pour effectuer le lien entre MATLAB® et ROS, le consortium TWO!EARS a décidé de poursuivre avec le développement et l'utilisation du *matlab-to-ros bridge*.

3.3.2.4 Navigation et localisation

La nature des scénarios de TWO!EARS implique que le robot doive pouvoir se mouvoir dans l'environnement de façon sûre et autonome. Il doit ainsi être capable de se situer précisément et de naviguer entre deux points de coordonnées (x, y, θ). Cela implique également la capacité de planification d'un chemin prenant en compte les contraintes cinématiques du robot, ainsi que l'aptitude à réagir à l'apparition d'obs-

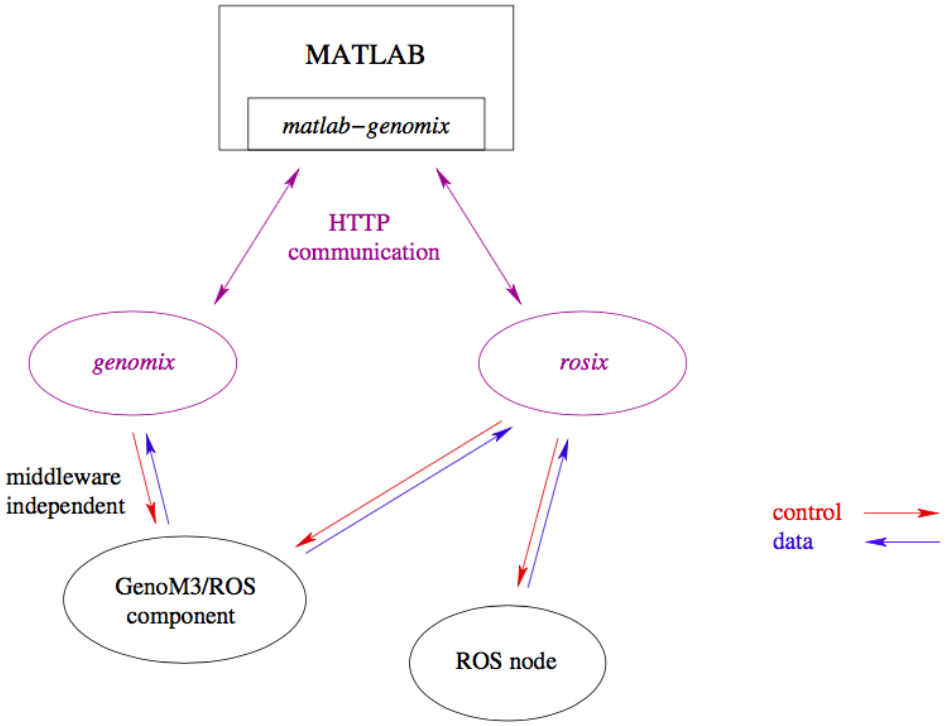


FIGURE 3.14 – MATLAB-TO-ROS BRIDGE — Utilisation de *genomix* et *rosix* afin de faire un pont entre ROS et MATLAB®. *genomix* permet de contrôler les composants *GenoM3* et de lire leurs flux de données, indépendemment du *middleware* employé. *rosix* peut contrôler et lire les données de n'importe quel nœud ROS de la couche fonctionnelle. *matlab-genomix* peut être un client de n'importe quel serveur *genomix* ou *rosix*.

tacles inattendus ou mobiles (des personnes par exemple). L'algorithme *Gmapping*[32] a été utlisé pour effectuer la cartographie de l'environnement selon une approche SLAM (cf. **Sec. 2.1.2.1**). Il s'agit d'un algorithme *open-source* utilisant un filtre à particule de type Rao-Blackwellized incluant les mesures LASER et odométriques, et les combinant avec le mouvement du robot. *Gmapping* a été implémenté en tant que paquet ROS[33] et inclus dans l'ensemble de paquets *ROS navigation*[34]. Cet algorithme requiert la définition précise d'un certain nombre de paramètres (sur un total de 36), comme la portée des capteurs LASER, le nombre de particules du filtre, la résolution de la carte etc.

La **Fig. 3.15** montre l'interaction entre les différents composants requis lors d'une tâche de navigation.

3.3.3 Discussion

Nous avons décrit dans cette section les caractéristiques matérielles aussi bien que logicielles de la plateforme robotique sur laquelle le logiciel TWO!EARS a été intégré. Un des points forts de la plateforme utilisée est le mannequin HATS doté

32. `http://openslam.org/gmapping.html`
33. `http://wiki.ros.org/gmapping`
34. `http://wiki.ros.org/navigation`

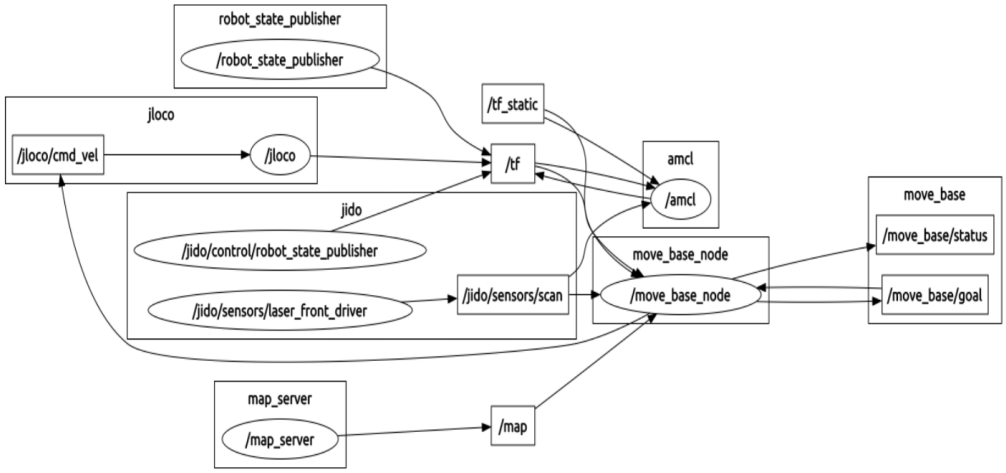


FIGURE 3.15 – Interaction entre les composants requis pour la navigation du robot. (*ellipses*) nœuds ROS ; (*boîtes*) topique ROS ; (*boîtes incluantes*) *namespaces* regroupant les ressources similaires.

d'une audition binaurale et auquel une vision stéréoscopique a été ajoutée par le LAAS, conférant au robot des caractéristiques humanoïdes.

Le logiciel TWO!EARS intégré à la platefore robotique va ensuite être évalué dans différents scénarios de l'état de l'art. La section suivante décrit ainsi les quatre principaux de scénarios utilisés pour les tests et la validation des différents composants du logiciel TWO!EARS, le modèle HTM étant un d'entre eux.

3.4 Scénarios de test

AFIN de valider l'ensemble des implémentations effectuées par le consortium TWO!EARS, des scénarios de test ont été mis au point. Ces scénarios sont de type *Dynamic Auditory Scene Analysis*[35] (DASA), le plus ambitieux étant celui de *Search & Rescue scenario*[36] (S&R). Les sections suivantes détaillent les quatre principaux scénarios utilisés pour l'évaluation du système TWO!EARS.

DASA-1

Titre : localisation de signaux de parole en condition multi-sources et reconnaissance de genre.

Description : une voix de femme est localisée en présence de quatre sources sonores de voix d'hommes réparties dans l'espace. Deux conditions différentes sont considérées : (i) les positions des sources masquantes sont connues *a priori*, (ii) les positions sont inconnues.

Tâches : (i) trouver la position de la voix cible, (ii) déterminer si une voix de femme est présente.

35. Analyse de scène audio dynamique.
36. Scénario de recherche et de sauvetage.

Validation : (i) erreur quadratique moyenne de l'azimut de la source cible, (ii) performance de la détection de la voix de femme, quantifiée par taux de bonne reconnaissance et taux de faux positifs.

DASA-2

Titre : reconnaissance de mots-clefs.

Description : reconnaissance de mots en présence de bruit et de réverbération. Les conditions de bruit incluent du bruit diffus en arrière-plan et jusqu'à quatre sources interférentes. La base de données *CHiMe challenge data* [261], contenant des enregistrements de sons domestiques dans un appartement (aspirateurs, enfants jouant, présence de musique etc.), a été utilisée pour cette évaluation. Ces enregistrements ont été superposés aux enregistrements binauraux des signaux de parole.

Tâches : identifier les mots-clefs.

Validation : taux d'erreur.

DASA-3

Titre : localisation et caractéristation de sources dans une pièce.

Description : il s'agit d'une version simplifiée du scénario de S&R dans lequel des sources sonores particulières d'intérêt (« victimes ») sont situées dans un appartement à une pièce. La détection des victimes implique leur localisation, la reconnaissance du locuteur (selon les trois groupes : homme, femme, enfant) et la reconnaissance de mots-clefs, dans des environnement acoustiques complexes. Des boucles de rétro-contrôles seront utilisées ici (comme le mouvement du robot et de sa tête par exemple). De plus, ce scénario nécessite la capacité de détecter des personnes visuellement.

Tâches : (i) identification de la position des sources sonores et orientation consécutive de la tête du robot vers elles, (ii) identification du genre des voix présentes dans la scène, (iii) classification des sources, (iv) identification des mots-clefs.

Validation : (i) erreur quadratique moyenne de l'azimut des sources sonores, (ii) taux de bonne reconnaissance du genre, (iii) taux de bonne classification, (iv) taux de détection de mots-clefs.

DASA-4

Titre : localisation et caractérisation de sources dans un environnement à plusieurs pièces.

Description : ce scénario correspond à une tâche de S&R dans un environnement multi-sources. L'environnement consiste en trois pièces communicantes dans lesquelles le robot peut bouger librement dans le but de détecter les sources d'intérêt (« victimes ») et éviter les sources dangereuses. La détection de victimes implique la reconnaissance homme/femme/enfant, la reconnaissance de mots-clefs dans un environnement acoustique complexe et la détection de signaux vocaux de type « cris » ou « pleurs ». Des boucles de rétro-contrôle seront également utilisées. La modalité visuelle sera ici également nécessaire.

Tâches : (i) identification de la position des sources sonores et orientation du robot vers elles, (ii) identification du genre des locuteurs, (iii) classification de sources incluant la capacité à discriminer les victimes des sources de danger, (iv) identification de mots-clefs, (v) détection des signaux vocaux de type « cris » ou « pleurs », et (vi) navigation dans l'environnement.

Validation : (i) erreur quadratique moyenne de l'azimut des sources sonores, (ii) taux de bonne reconnaissance du genre, (iii) taux de bonne classification, (iv) taux de détection de mots-clefs, (v) capacité à détecter les signaux vocaux de type « cris » ou « pleurs » (taux de succès et taux de faux positifs), et (vi) mesure du temps pris pour la navigation dans l'environnement pour la localisation de chaque source cible.

3.4.1 Discussion

Le modèle HTM sera testé en conditions simulées et réelles sur les scénarios DASA-3 et DASA-4. Certains éléments de validation sont en revanche en-dehors du domaine de notre modèle, notamment la reconnaissance de mots-clefs ou la reconnaissance du genre d'un locuteur.

3.5 Conclusion du Chapitre

Ce chapitre a été dédié à une description du projet Two!Ears, projet ayant constitué un cadre dans lequel le modèle HTM a été implémenté. Two!Ears a été un projet européen ambitieux et innovant dont le but a été de créer des modèles de l'audition binaurale incluant une dimension *active*, *via* notamment l'utlisation de mouvements de tête, et intégrant la multimodalité grâce à la vision. Ayant pour but l'intégration de ces modèles dans une plateforme robotique mobile, l'approche générale adoptée au sein du projet s'est toujours située entre la forte inspiration biologique d'un côté et les contraintes de la robotique d'un autre côté. Malgré la difficulté de créer une architecture transversale, allant de l'aquisition des signaux à l'émergence de phénomènes cognitifs en passant par les problèmes de navigation robotique dans un environnement, un logiciel *open-source*, complet et riche a pu être développé.

Une des forces de ce logiciel est qu'il s'appuie sur l'utilisation intensive d'un Blackboard, structure permettant une communication facile et homogène entre différentes entités du système global. Ces entités sont les *Knowledge Sources*, chacune dédiée à une analyse précise et restreinte des données perçues par le robot. Tous ces experts travaillent en commun pour résoudre le « problème » suivant : comment comprendre un environnement audiovisuel de façon performante et selon une approche bioinspirée ?

En effet, la définition même du projet Two!Ears a été formalisée par la phrase suivante : « *Read the world with Two !Ears* ». L'audition binaurale, similaire à l'Homme, est une contrainte autant qu'un atout : elle complique grandement les tâches de localisation et d'identification par diminution du nombre de capteurs mais permet de

s'inspirer des excellentes performances observées et mesurées chez l'humain. D'autre part, l'ajout d'une vision binoculaire est bien plus que l'apposition d'une paire de caméras : elle permet de mettre à contribution une autre source de connaissances pouvant éventuellement être utilisée afin de lever d'éventuelles ambiguïtés sur l'analyse de la scène audio. Comme nous l'avons mis en avant dans le **Chap. 2**, la multimodalité est un énorme atout de la perception animale. Notre modèle sera d'ailleurs le seul a utiliser la vision.

Le modèle HTM a essentiellement fait partie du WP4, dédié formellement à l'identification puis à l'implémentation de boucles de rétro-contrôles permettant de donner au robot des prémisses de cognition. Se basant sur les connaissances générées par les différents experts, notre modèle tentera de les rassembler et d'en tirer du *sens*. Notre travail a également fait partie du WP5, dédié au travail sur les robots Jido et Odi, mais l'équipe du LAAS a été le principal contributeur des achèvements atteints dans ce WP.

Le chapitre suivant consiste justement en l'introduction du modèle HTM, maintenant que tous les pré-requis ont été décrits en détail.

Chapitre 4

Introduction du modèle HTM

Nous avons posé toutes les bases conceptuelles et théoriques sur lesquelles le modèle HTM se base au **Chap. 2** puis nous avons décrit en détail le projet Two!Ears dans lequel le modèle se situe au **Chap. 3**. Il est temps désormais d'introduire le modèle en lui-même. Le modèle HTM est composé de deux modules, chacun d'entre eux se basant sur des définitions et notations communes. Ainsi, ce chapitre a pour but de présenter les bases théoriques sur lesquelles ils se basent mais aussi l'ensemble des notations qui seront utilisées lors des évaluations de ces deux modules.

Le modèle Head Turning Modulation est un module **attentionnel bas-niveau** ayant pour but de **moduler les mouvements de tête d'un robot** au cours de l'exploration d'un environnement **inconnu**. La modulation de ces mouvements de tête englobe leur génération aussi bien que leur inhibition. Le modèle HTM est composé de deux principaux modules :

- le module *Dynamic Weighting*[1] [262, 263] (module DW, **Chap. 5**) dont le rôle est de comprendre l'environnement en cours d'exploration à travers la notion de *Congruence* qui peut être décrite, dans le contexte du modèle HTM, comme une mesure de l'*importance* d'un objet audiovisuel.

- le module *Multimodal Fusion & Inference*[2] [225] (module MFI, **Chap. 6**) dont le rôle est d'apprendre la relation entre les différentes modalités qui caractérisent un *objet*, c'est-à-dire, dans le champ d'application du projet Two!Ears, les modalités audio et visuelles. De plus, le module MFI est capable, à partir de cet apprentissage, d'effectuer une inférence d'une modalité éventuellement manquante, comme lors d'une occlusion visuelle par exemple.

Ces deux modules aboutissent à la détermination de l'objet présent dans l'environnement nécessitant l'attention du robot. Par *attention*, nous entendons la zone spatiale vers laquelle le robot devrait tourner sa tête, c'est-à-dire la zone sur laquelle il lui serait nécessaire ou intéressant de focaliser ses capteurs dans le but de récupérer le plus d'information possible sur cet objet.

1. module de Pondération Dynamique.
2. module de Fusion et d'Inférence Multimodale.

Le modèle Head Turning Modulation a été implémenté en tant que KS (nommée ainsi HeadTurningModulationKS, cf. **Sec. 7.2.1**) à l'instar de tous les autres sources de connaissance et experts du projet Two!Ears. Cette implémentation favorise grandement la communication entre les KS ainsi que le regroupement du résultat de leurs analyses, *via* l'architecture de type *Blackboard*. En effet, le modèle Head Turning Modulation traite des données aussi bien audio que visuelles afin de faire émerger la notion d'« objet » au sein de la représentation interne du robot de l'environnement qu'il explore. Il est donc nécessaire d'avoir accès aussi facilement et rapidement que possible à toutes les données nécessaires au module DW et au module MFI.

La **Sec. 4.1** exposera les définitions et les notations primordiales au développement du modèle.

La **Sec. 4.2** approfondira la notion d'objet multimodal telle qu'elle a été comprise au sein du modèle HTM, notamment sur la base des travaux de recherche exposés à la **Sec. 2.2** et à la **Sec. 2.3**.

La **Sec. 4.3** présentera l'environnement de simulation créé pour évaluer le modèle HTM lorsque les KS du système Two!Ears étaient encore en cours de développement et que le robot n'était pas encore pleinement fonctionnel.

La **Sec. 4.4** enfin détaillera les critères d'évaluation utilisés pour juger des performances du modèle.

4.1 Définitions & Notations

Le modèle HTM est un modèle computationnel d'exploration sémantique d'un environnement inconnu incluant un mécanisme attentionnel et un processus d'intégration multimodale des informations audio et visuelles. Une des bases conceptuelles du modèle est la notion d'*Objet* multimodal. Nous avons choisi de mettre cette notion au centre de la représentation interne de l'environnement que le robot tente de se construire : un environnement en cours d'exploration est défini par les objets qui le composent. Ainsi, de cette notion d'objet découle la notion d'*Environnement* dans lequel il est inclus ainsi que la notion de *représentation interne* de ces environnements. Toutes ces notions doivent donc être précisément définies et formalisées par une notation précise et adéquate. Cette section présente ainsi les définitions et les notations ayant servies de base à la création du modèle HTM.

4.1.1 Définitions

Tout d'abord, l'ensemble du modèle HTM concerne la création d'un algorithme dotant un robot d'une perception multimodale capable de générer un comportement attentionnel motivé par l'exploration d'un environnement inconnu aboutissant à une représentation interne du monde qu'il explore. Nous définissons ainsi le Robot comme suit :

Définition 1. *Un* ROBOT *est défini comme une entité mobile dotée de capteurs, capable de traiter les données qu'il perçoit grâce à ceux-ci et de réagir en fonction du résultat de ces analyses.*

Les données que le ROBOT perçoit sont particulièrement celles captées par les caméras et les microphones dont il est doté (cf. **Sec. 3.3**). Ces données proviennent des sources audiovisuelles présentes dans l'environnement.

Définition 2. *Une* SOURCE *est définie comme une entité réelle de l'*ENVIRONNEMENT *appartenant à une catégorie audiovisuelle et située à une position spatiale.*

Avant que ces SOURCES ne soient intégrées par le robot dans sa représentation interne de l'ENVIRONNEMENT, les signaux qu'elles émettent sont d'abord considérés comme des événements audiovisuels ne possédant pas de cohérence temporelle ni de sens. Nous définissons ainsi un ÉVÉNEMENT comme suit :

Définition 3. *Un* ÉVÉNEMENT *est défini comme l'apparition d'un stimulus audio, visuel ou audiovisuel émis par une* SOURCE *dans un* ENVIRONNEMENT *donné.*

Le but du modèle est de permettre au ROBOT de comprendre ces événements comme des OBJETS audiovisuels. Nous définissons ainsi un OBJET commme suit :

Définition 4. *Un* OBJET *est la représentation interne du* ROBOT *d'une* SOURCE. *Un* OBJET *est caractérisé par des labels audio et visuels ainsi qu'une position dans l'espace. Il rassemble également l'ensemble des* ÉVÉNEMENTS *perçus ayant été associés à la* SOURCE *qu'il représente.*

L'apparition d'une SOURCE dans un ENVIRONNEMENT, ainsi que les caractéristiques de celle-ci, sont un moyen de définir cet ENVIRONNEMENT, que nous définissons ainsi :

Définition 5. *Un* ENVIRONNEMENT *est défini comme une entité spatiale caractérisée par l'ensemble des* OBJETS *qui y sont présents.*

Une fois les ÉVÉNEMENTS associés à des SOURCES et interprétés comme des OBJETS au sein de la représentation interne de l'environnement en cours d'exploration, le système va tenter de leur assigner du sens par analyse de leur Congruence aux autres objets (détaillé au **Chap. 5**) dans cet environnement. En fonction de cette mesure, un mouvement de tête sera éventuellement généré dans leur direction. Ainsi, nous ajoutons la définition de la *Focalisation* :

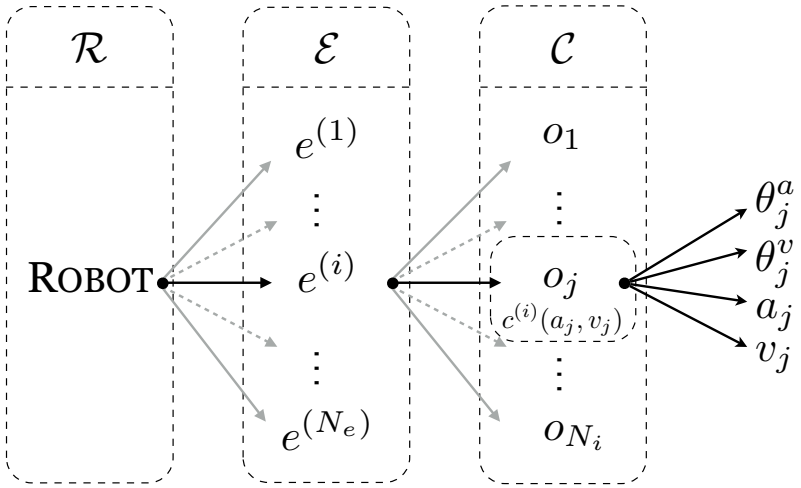


FIGURE 4.1 – ORGANISATION SCHÉMATIQUE DU MODÈLE HTM — Un robot $\mathcal{R}$ est dans un environnement $e^{(i)} \in \mathcal{E}$ composé de plusieurs objets $o_j \in \mathcal{C}$ caractérisé par leur position $[\theta_j^a, \theta_j^v]$ et leurs labels audio et visuel $[a_j, v_j]$ (cf. **Sec. 4.1**).

Définition 6. *Un* OBJET *est dit « focalisé » lorsque la tête du robot lui fait face en réponse à une requête du module DW ou du module MFI.*

Pour récapituler : une SOURCE est une entité audiovisuelle de l'ENVIRONNEMENT, émettant des EVÉNEMENTS audiovisuels que le ROBOT tentera d'inclure dans sa représentation interne en tant qu'OBJET par génération de mouvements de tête (FOCALISATION).

Enfin, nous introduisons dès à présent la notion de « Robot naïf » $\mathfrak{R}^{\mathfrak{n}}$ qui nous servira durant toutes les évaluations à des fins de comparaison et dont les différentes caractéristiques seront détaillées au cours de ce chapitre, notamment en terme de comportement attentionnel et de capacités d'intégration multimodale. Ce robot a été inspiré des travaux de recherche de BENOÎT GIRARD *et al.* [264] en 2002, publiés au sein de la revue *From Animals to Animats*, et ayant développé un algorithme d'exploration intégré à une plateforme robotique, similaire à celui de MEYER *et al.* [51] décrit à la **Sec. 2.1.2**. L'algorithme développé par GIRARD *et al.* est comparé à un algorithme de type *Winner-Takes-All* ne réalisant aucune intégration des différentes données à disposition et se contentant de ne sélectionner que l'action motrice la plus attractive au temps t. De façon similaire, notre robot naïf consistera en une implémentation d'algorithme prenant des décisions sur la base d'informations à court-terme et ne pouvant ainsi réaliser d'intégration temporelle ou multimodale performante.

4.1.2 Notations

Cette section décrit l'ensemble des notations utilisées dans le cadre de la formalisation du modèle HTM. Ces notations font écho aux définitions présentées ci-dessus

et seront utilisées dans toute la suite.

Soit $\mathcal{R}$ le robot et $\mathcal{E}$ un environnement qu'il explore avec

$$\mathcal{E} = \{e^{(1)}, e^{(2)}, \ldots, e^{(N_e)}\} \tag{4.1}$$

où $e^{(i)} \in \mathcal{E}$ représente le i-ème environnement exploré par $\mathcal{R}$, et N_e le nombre d'environnements déjà explorés. Chaque environnement $e^{(i)}$ est défini comme un ensemble d'objets o_j tel que

$$e^{(i)} = \{o_1, o_2, \ldots, o_{N_i}\} \tag{4.2}$$

où N_i est le nombre d'objets détectés dans l'environnement $e^{(i)}$. Chaque objet o_j est défini par ses angles audio et visuel $[\theta_j^a, \theta_j^v]$ relatifs au robot, ainsi que par des labels audio et visuels $[a_j, v_j]$, de telle sorte que

$$o_j = \{\theta_j^a, \theta_j^v, a_j, v_j\} \tag{4.3}$$

Nous introduisons également la formalisation de la notion d'Evénement (cf. **Déf. 3**) différente de celle d'Objet en cela qu'il correspond à la réalité terrain susceptible d'être perçue en partie ou en entier par le robot, et dont la perception est sujette à erreurs, que ce soit dans la classification ou la localisation. Un Evénement ψ_k, de façon similaire à un objet o_j, est également caractérisé par ses angles audio et visuels $[\theta_k^a, \theta_k^v]$, ainsi que par des labels audio et visuels $[a_k, v_k]$, tel que :

$$\psi_k = \{\theta_k^a, \theta_k^v, a_k, v_k\} \tag{4.4}$$

Un des buts du modèle HTM, et également une façon de juger la performance du modèle, est d'être capable de transformer chaque Evénement ψ_k de l'environnement $e^{(i)}$ en objet o_j avec :

$$\psi_k = \{\theta_k^a, \theta_k^v, a_k, v_k\} \rightarrow o_j = \{\theta_j^a, \theta_j^v, a_j, v_j\} \tag{4.5}$$

Cette équivalence peut être obtenue soit par accès direct à toutes les informations de l'événement avec éventuelle correction des erreurs de classification par le module MFI ; soit par inférence des informations manquantes, toujours par le module MFI, si le robot n'a pas accès à toutes les données de l'événement et s'il est capable d'inférer correctement ces données manquantes. Un événement pouvant ne pas émettre de son ou ne pas être visible, nous précisons l'**Eq. 4.5** :

> $\ddot{\psi}_k$, si l'Evénement est audio et visuel, donc que toutes les modalités sont disponibles pour le robot ;
> $\dot{\psi}_k$, si l'Evénement est audio ou visuel, donc qu'une des modalités est manquante.

Chaque Evénement est produit par une Source $\mathcal{S}_k$ également définie par la catégorie audiovisuelle à laquelle elle appartient. Ainsi, pour résumer : un événement

audiovisuel ψ_k survient dans l'environnement $\varepsilon^{(i)}$ lorsqu'une source $\mathcal{S}_k$ se met à émettre du son (l'émission de données visuelles est constante dans le temps), événement que le modèle HTM cherchera à transformer en une représentation en objet o_j stable dans le temps et robuste aux éventuelles erreurs de classification des experts dédiés.

Définissons maintenant les *catégories audiovisuelles* $c^{(i)}(a, v)$ du i-ème environnement par :

$$c^{(i)}(a, v) = \{o_j \in e^{(i)}, a_j = a, v_j = v\} \tag{4.6}$$

où $c^{(i)}(a, v)$ représente la collection d'objets partageant les mêmes labels audio et visuels a et v respectivement. Toutes les catégories du i-ème environnement sont rassemblées dans un ensemble de catégories $\mathcal{C}^{(i)}$ tel que $\mathcal{C}^{(i)} = \{c^{(i)}(a, v)\}$.

Le champ de vision du robot est une caractéristique prise en compte dans les algorithmes de localisation visuelle absolue ainsi que dans la détection et la reconnaissance d'objets. Elle est notée θ_{fov} (avec *fov* signifiant *field of view*).

Au début de chaque expérimentation, le robot est placée à sa position dite de *repos*, définissant ainsi le repère à $\theta_0 = 0°$ utile pour les algorithmes de localisation ainsi que pour le calcul des ordres moteurs générés par les deux modules du modèle HTM. Cette position sera notamment utilisée lorsqu'aucun objet n'est présent dans l'environnement ou lorsqu'aucun objet ne nécessite l'attention du robot.

D'autre part, nous définissons $\theta_{\text{DW}} \in [0, 359]$ et $\theta_{\text{MFI}} \in [0, 359]$ comme les angles de rotation de la tête du robot que le module DW et le module MFI respectivement sont susceptibles de requérir (cf. **Sec. 5.3.2** & **Sec. 6.4** respectivement).

Enfin, nous définissons t comme le temps discret mesuré en itération, que ce soit lors d'expérimentations en simulation ou en environnement réel. En effet, le *Blackboard* assure que toutes les KS, bien que déclenchées de façon séquentielle, aient toujours le même indice temporel comme repère. Le temps du *Blackboard* est, lui, mesuré en millisecondes. Lors d'une requête de résultats de n'importe quelle KS, le temps récupéré sera le même pour toutes. Mais par souci de simplicité, et sachant que le modèle HTM ne se base pas sur une notion de temps *véritable* directement, mais plutôt sur un nombre de trames ou d'itérations, nous avons décidé de définir le temps t comme un temps discret.

4.2 La notion d'objet

Nous avons déjà défini plus haut la notion d'Objet (cf. **Déf. 4**) Cependant, il est nécessaire de préciser cette notion car elle va conditionner l'ensemble du modèle HTM. Au sein du modèle, un Objet est la fusion des données audio et visuelles d'identité et de localisation obtenues depuis les experts Two!Ears dédiés. L'approche basée sur la notion d'objet permet d'émettre des hypothèses sur l'organisation des données perçues, notamment sur le plan temporel. Les mécanismes neuronaux, et plus généralement cérébraux, exposés à la **Sec. 2.2** dédiée à

la perception auditive et visuelle, et à la **Sec. 2.3** dédiée aux mécanismes attentionnels, montrent que le cerveau se base amplement sur une représentation multimodale des événements sensoriels apparaissant dans son environnement, représentation possèdant une caractéristique profondément temporelle. C'est ainsi que les aires sensorielles possèdent des capacités de prédiction (cf. **Sec. 2.3.1.4**). De cette représentation et des capacités prédictives résultantes, des réactions comportementales attentionnelles sont rendues possibles. L'imprédictibilité d'un changement d'une ou plusieurs caractéristiques d'un stimulus audio ou visuel peut aboutir à la génération d'un mouvement de tête par exemple. A partir de cette base, nous posons donc l'hypothèse suivante :

Hypothèse 1. Cohérence de l'identité : *Un objet possède une certaine cohérence temporelle en terme d'identité.*

Ainsi, nous considérons qu'un objet appartient à une même classe audio et visuelle au cours d'un certain temps. Cette cohérence peut également être comprise comme suit : si au temps t, un objet appartient à une classe audiovisuelle donnée, sa classe au temps $t+1$ sera strictement la même. Cette propriété nous permet d'effectuer une intégration temporelle des données issues des classifieurs. En effet, une fois la position spatiale d'un objet audiovisuel déterminée, toutes les informations des experts d'identification audio et visuels pourront être rassemblées au sein du même Objet. Cette intégration consistera en plus qu'une simple concaténation des données ne permettant pas de profiter du nombre grandissant des données aquises — l'« expérience » — au cours de l'émission de stimuli audio et visuels d'un objet donné, mais sera la fusion de toutes les données perçues, modalité par modalité, pour un objet donné, selon :

$$\mathbf{P}(o_j)[t] = \frac{1}{N_t} \sum_{n=t_i}^{n=t} \mathbf{P}(o_j)[n] \tag{4.7}$$

avec N_t le nombre de trames temporelles pour lesquelles des données ont été assignées à l'objet o_j. Cependant, la durée temps d'intégration ne peut pas être trop longue car elle pourrait alors empêcher un objet audiovisuel de changer de catégorie au cours du temps : un homme qui *crie* ne pourrait pas se mettre à être un homme qui *parle*. Ainsi, la durée d'intégration, exprimée par le terme N_t, a été réduite à 10 trames, soit environ 5 s de temps réel. Cette étape d'intégration temporelle permettra ainsi d'initier la fusion intramodale de classifieurs ainsi que d'améliorer l'apprentissage du lien entre les différentes modalités constitutives d'un objet. Bien que simple, le moyennage des données issues des experts repose sur le concept d'objet perceptuel aux bases cérébrales fortes. Le vecteur $\mathbf{P}(o_j)[t]$ est le vecteur qui sera envoyé au module MFI, décrit plus tard, afin (i) qu'une étape d'apprentissage puisse être effectuée et (ii) qu'une éventuelle correction d'erreur puisse être effectuée par fusion multimodale des données. Ainsi, l'assignation à un objet o_j d'une catégorie audiovisuelle sera faite après analyse du module MFI du vecteur $\mathbf{P}(o_j)[t]$.

D'autre part, nous nous plaçons dans des conditions de sources statiques, contrainte imposée par les scénarios de Two!Ears. En effet, les scénarios de tests présentés

au **Sec. 3.4** ne contiennent que des sources statiques. Dans la majeure partie du développement du modèle HTM, l'accent a été mis sur la validation des mouvements de tête vers des sources d'intérêt. Ainsi, les capacités mobiles du robot, au sens de mouvement du corps entier du robot, n'ont pas été prises en compte. Cela entraine ainsi également une cohérence de localisation : un objet possède toujours une même localisation spatiale, objet de l'hypothèse suivante :

Hypothèse 2. COHÉRENCE DE LA POSITION : *Un objet possède une cohérence temporelle en terme de localisation spatiale.*

4.3 Environnement de simulation : HtmTestBed

AFIN de tester les différents composants du logiciel TWO!EARS en attendant que le robot soit pleinement fonctionnel, plusieurs environnements de simulations ont été développés. Un environnement de simulation a été développé au sein de TWO!EARS permettant d'émuler des environnements acoustiques ainsi que le robot, possiblement mobile. Cependant, le modèle HTM nécessite quasiment l'ensemble des KS du logiciel pleinement fonctionnelles ainsi que l'intégration de la modalité visuelle. Or, cette dernière n'étant pas prise en compte dans le système de simulation de TWO!EARS, il nous a été nécessaire de créer notre propre système. Cette section détaille le système HtmTestBed (HTMtb) permettant de simuler des environnements audiovisuels, le robot et ses mouvements de tête, ainsi que les KS nécessaires au fonctionnement du modèle HTM.

Afin de pouvoir tester le modèle HTM il a été nécessaire de simuler les KS sur lesquelles il se base, nomément : AUDITORYIDENTITYKS, DNNLOCATIONKS, VISUALIDENTITYKS et la VISUALLOCATIONKS. Sachant que le modèle HTM se base exclusivement sur la sortie de ces experts, seulement la simulation des vecteurs qu'ils produisent a due être effectuée (cf. **Fig. 4.2**) :

(a) Localisation audio : $\mathbf{\Theta}^a[t] = (\theta_1^a[t], \ldots, \theta_N^a[t])^T \mod 360$ (cf. **Eq. 3.5**) ;

(b) Identification audio : $\mathbf{P}^a[t] = (p_1^a[t], \ldots, p_{N_a}^a[t])^T$ (cf. **Eq. 3.6**) ;

(c) Localisation visuelle : $\Theta^v[t] = \theta^v[t] + \theta_{head} + \theta_{torso} \mod 360$ (cf. **Eq. 3.13**) ;

(d) Identification visuelle : $\mathbf{P}^v[t] = (p_1^v[t], \ldots, p_{N_v}^v[t])^T$ (cf. **Eq. 3.7**).

Comme détaillé à la **Sec. 3.2.3**, il y a un expert d'identification/détection par classe audio, idem pour la modalité visuelle. Chaque sortie individuelle de ces experts est ensuite concaténée en un vecteur de probabilités *a posteriori*. Afin de tester le modèle HTM, un ensemble d'experts d'identification et de localisation a été implémenté permettant de reproduire le comportement des véritables experts de TWO!EARS. En tout, 38 experts audio et 19 experts visuels ont été créés, avec un total de 722 combinaisons possibles.

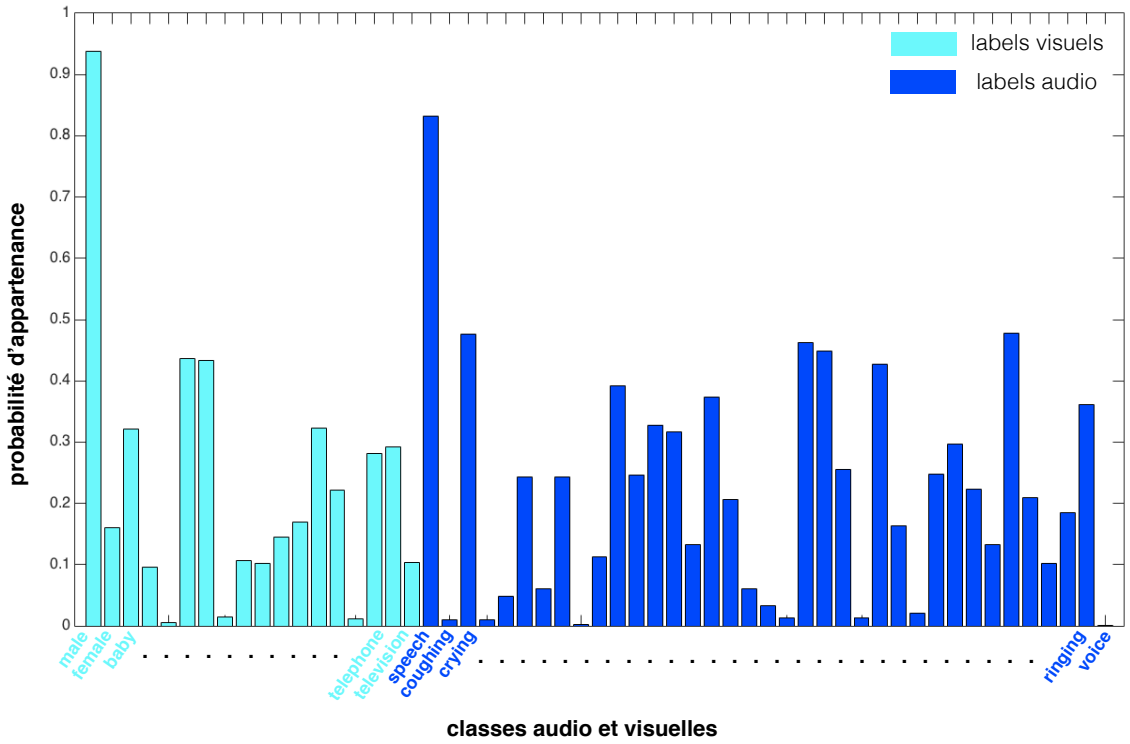


FIGURE 4.2 – EXEMPLE DE DONNÉES SIMULÉES — Une source audiovisuelle simulée par le HtmTestBed émet des données sonores et visuelles interprétées par les experts d'identification audiovisuels simulés. Chaque expert émet une probabilité d'appartenance à la catégorie pour laquelle il a été spécifiquement entrainé. L'ensemble des probabilités sont rassemblées en un vecteur unique, illustré ici par cet histogramme. La catégorie audiovisuelle réelle simulée ici correspondrait à *male speech*. Un mauvais appariement audiovisuel généré par le simulateur se traduirait par une composante maximum audio ou visuelle erronée.

Les experts d'identification et de localisation de TWO!EARS pouvant produire des données erronées — d'autant plus que les données sont complexes, comme dans le cas multisource — les données générées par le HtmTestBed l'ont été avec un certain taux d'erreur rendant compte d'un comportement proche de celui des véritables experts. Un taux d'erreur des experts d'identification audio $\varepsilon_{\mathcal{P}^a}$ ainsi qu'un taux d'erreur des experts d'identification visuels $\varepsilon_{\mathcal{P}^v}$ ont été définis. Par exemple, $\varepsilon_{\mathcal{P}^a} = 35\%$ signifie qu'à chaque trame, il y a une probabilité de 35% qu'une erreur de classification audio survienne. Cela se traduira par une composante maximale du vecteur $\mathbf{P}^a[t]$ fausse, étant donnée la classe réelle $\mathcal{C}^a[t]$ de l'objet o_j. De façon similaire, un taux d'erreur des experts de localisation a été défini $\varepsilon_{\mathbf{\Theta}^a}$ & $\varepsilon_{\mathbf{\Theta}^v}$. Son fonctionnement est le même que les taux d'erreur pour les experts d'identification.

Au cours des simulations, un indice temporel t est utilisé. Cet indice temporel correspond à une trame, de longueur égale à 500 ms, en accord avec les trames employées par les experts de localisation et d'identification. Ainsi, une simulation de 1000 pas de temps correspondra à une expérience ayant duré 500 secondes ($\approx 8,3$ minutes).

Dans les différentes expériences effectuées avec cet environnement de simulation, entre 1 et 10 sources audiovisuelles ont été simulées. De chacune de ces sources, un événement audiovisuel peut survenir dont la durée varie de 15 à 35 trames (7, 5 s à

17, 5 s en temps du Blackboard), l'émission d'un nouvel événement ne pouvant survenir avant une période minimale de 15 trames (7, 5 s) et maximale de 50 trames (25 s). Ces durées — relativement longues comparées à des durées réalistes — ont été choisies en prenant en compte la durée des événements sonores utilisées lors des expériences menées sur le vrai robot. Ces durées varieront entre 15 et 20 s : même si une itération du Blackboard dure théoriquement 500 ms, nous verrons que ce temps peut parfois augmenter jusqu'à 2 à 3 s et que la répétition d'un stimulus sonore d'une même classe est nécessaire pour que les experts Two!Ears convergent vers une classification correcte (même si non exempte d'erreur après ces répétitions). De plus, cette durée durant laquelle les événements simulés émettent prennent également en compte le temps que le modèle HTM, en conditions réelles, met à générer des mouvements de tête : celui-ci nécessite parfois quelques trames. Ainsi, pour éviter que le robot ne tourne sa tête de façon incessante, nous avons choisi une durée qui, même si elle semble relativement longue, est cohérente avec le comportement du système Two!Ears et du modèle HTM en conditions réelles. De ces événements, les experts de localisation et d'identification simulés seront en charge de produire une probabilité d'appartenance à une des classes audio ou visuelle possible, ainsi qu'une position en azimut. Un point important de cet environnement spécifique à l'évaluation des composants du modèle HTM : le robot est immobile. En effet, le modèle étant dédié à la modulation des mouvements de tête du robot, seule cette capacité motrice a été incluse dans le système de simulation.

D'autre part, afin de caractériser les différents scénarios créés pour la validation du modèle, nous introduisons quelques notations supplémentaires :

- $n_{\mathcal{S}}$ est défini comme le nombre de sources audiovisuelles présentes dans l'environnement ;
- n_{sim}^{max} est défini comme le nombre maximum autorisé de sources émettant du son simultanément ;
- T est le nombre de pas de temps d'une simulation (un pas de temps étant censé correspondre à une trame temporelle définie par le *Blackboard* de 500 ms).

Ainsi, deux simulations générées avec ces mêmes paramètres auront des décours temporels certainement différents puisqu'aucun contrôle de ce décours n'est ici proposé. Cette absence de contrôle par défaut (puisqu'il est toujours possible d'imposer un scénario temporel) a été faite volontairement : nous avons souhaité, à chaque création d'un scénario, (i) générer l'assignation des catégories audiovisuelles aux sources présentes dans l'environnement de façon aléatoire et (ii) que les événements audiovisuels générés par chaque source apparaissent également de façon aléatoire.

La **Fig. 4.3** illustre le système HTMtb créé pour tester le modèle HTM.

(a) cadre, haut—gauche : environnement contenant le robot ainsi que les sources audiovisuelles simulées ;

(b) graphe, haut—milieu : taux de classification audiovisuelle de trois algorithmes de correction et de fusion des données issues des experts ;

(c) graphe, milieu—milieu : (*gris clair*) illustration de la détermination de l'objet désigné comme cible de l'attention du robot ; (*gris foncé*) origine de l'ordre moteur : module DW ou module MFI ;

Catégories audiovisuelles $\mathcal{C}$			
1	*male* speech	26	*door* closing
2	*male* coughing	27	*siren* beeping
3	*male* crying	28	*siren* alert
4	*male* eating	29	*train* alert
5	*male* piano	30	*train* braking
6	*male* violin	31	*train* accelerating
7	*male* laughing	32	*car* braking
8	*female* speech	33	*car* accelerating
9	*female* crying	34	*car* drifting
10	*female* eating	35	*glass* breaking
11	*female* singing	36	*glass* clinking
12	*female* laughing	37	*bed* squeaking
13	*female* flute	38	*wind* blowing
14	*female* screaming	39	*rain* falling
15	*female* harp	40	*switch* switched
16	*baby* crying	41	*loudspeaker* music
17	*baby* laughing	42	*loudspeaker* radio
18	*baby* screaming	43	*television* movie
19	*bird* whistling	44	*television* news
20	*bird* flying	45	*television* videogame
21	*dog* barking	46	*telephone* ringing
22	*dog* panting	47	*telephone* voice
23	*cat* meaowing	48	*telephone* music
24	*cat* scratching	49	*thunder* striking
25	*door* knock		

TABLE 4.1 – Classes audiovisuelles utilisées pour les simulations.

(d) graphe, bas—milieu : nombre de mouvements de tête vers chacune des sources audiovisuelles simulées dans l'environnement. (*bleu*) mouvements de tête générés par le modèle HTM ; (*rouge*) mouvements de tête générés par un robot naïf ;

(e) histogramme, haut—droite : taux de classification audiovisuelle au temps t par les trois algorithmes de correction et de fusion des données issues des experts ;

(f) histogramme, milieu—droite : ratio entre le nombre de mouvements de tête générés par le module MFI et le module DW ;

(g) histogramme, bas—droite : nombre de mouvements de tête générés par le modèle HTM comparé au robot naïf $\mathfrak{R}_n$.

Le **Tab. 4.1** liste toutes les classes audio, visuelles et les combinaisons audiovisuelles utilisées dans le système HTMtb.

D'autre part, le modèle HTM étant composé de deux modules différents, il a été nécessaire de les tester séparément. Pour cela, nous avons conçu la HEADTURNINGMODULATIONKS de façon à pouvoir utiliser sélectivement le ou les modules souhaités.

Enfin, l'ordinateur sur lequel toutes les simulations ont été effectuées est doté d'un processeur Intel core i7 cadencé à $1,7$ GHz et 8Go de mémoire cadencée à 1600 MHz

de type DDR3.

4.4 Critères d'évaluation du modèle

Afin d'évaluer le modèle HTM, et ses deux modules séparément, plusieurs critères ont dû être mis en place et implémentés. La littérature est fournie quant à chacune des parties du modèle : apprentissage par un réseau de type SOM, fusion de classifieurs, motivations à explorer etc. Cependant, très peu de modèles intègrent l'ensemble de ces composants dans un contexte aux nombreuses contraintes. Ainsi, il a fallu déterminer nos propres critères, fortement inspirés de ceux de la littérature, pour comparer les résultats obtenus grâce au modèle HTM à d'autres types de comportements.

4.4.1 Mouvements de tête

Pour comparer la capacité du modèle à déclencher puis à inhiber les mouvements de tête du robot, un comportement « naïf » a été implémenté : à chaque fois qu'un nouvel événement apparaît dans l'environnement, un mouvement de tête sera généré. Ce comportement se rapporte à la motivation par la Nouveauté (cf. **Sec. 2.1.3**) en cela que la seule caractéristique prise en compte est la détection d'un changement d'activité d'une source audiovisuelle et l'attraction que ce changement provoque. Cette attraction peut également être comprise comme l'attraction par la saillance d'un événement audio apparaissant au sein d'un environnement acoustique devenu constant, du point de vue informationnel (de par le temps durant lequel les sources actuelles ont émis). Aucune forme de contextualisation de l'événement n'est prise en compte par ce robot, son contenu sémantique non plus. Ce comportement du robot $\mathfrak{R}^{n}$ est défini selon la propriété suivante :

Propriété 1. *Les mouvements de tête du robot $\mathfrak{R}^{n}$ ne sont dépendants que de l'apparition d'événements ψ dans l'environnement $e^{(i)}$ en cours d'exploration.*

Ainsi :

1. Aucun contenu sémantique n'est pris en compte.
2. Aucune contextualisation n'est prise en compte.

4.4.2 Fusion de classifieurs

Nous avons introduit à la **Sec. 2.4.2** la notion de *Decision Support System* (DSS) grâce auquel le modèle HTM effectuera une fusion des experts d'identification (les « classifieurs ») présentés à la **Sec. 3.2.3**. Le DSS que nous avons implémenté se base principalement sur l'architecture du M-SOM au sein du module *Multimodal Fusion & Inference*, présenté au chapitre **Chap. 6**. L'ensemble du modèle HTM effectuera deux types de fusion : une fusion intramodale et une fusion intermodale. La première consiste en la fusion des données au sein d'une même modalité tandis

que la seconde concernera la fusion entre les modalités. Nous présentons ici les définitions et notations utilisés durant cette étape de fusion.

Soit $\mathcal{D}^{\mathbf{a}} = \{\mathcal{D}^a_1, \ldots, \mathcal{D}^a_k, \ldots, \mathcal{D}^a_{N^a}\}$ l'ensemble des N^a classifieurs dédiés à l'identification audio (cf. **Eq. 3.6**). De façon similaire, soit $\mathcal{D}^{\mathbf{v}} = \{\mathcal{D}^v_1, \ldots, \mathcal{D}^v_l, \ldots, \mathcal{D}^v_{N^v}\}$ l'ensemble des N^vclassifieurs dédiés à l'identification visuelle (cf. **Eq. 3.7**). Chaque classifieur $\mathcal{D}^a_k$ et $\mathcal{D}^v_l$ au temps t produit une probabilité $p^{a,v} \in \mathbb{R}^+ = [0, 1]$ sur l'appartenance à une classe audio et visuelle des vecteurs d'entrée audio et visuel respectivement. Toutes ces probabilités sont indépendantes entre elles et sont qualifiée de *fuzzy outputs* en cela qu'elle ne sont qu'une *probabilité* d'appartenance à une catégorie et non le reflet d'une décision certaine.

Définissons maintenant $d^a[t]$ et $d^v[t]$ comme la décision binaire prise au temps t sur l'appartenance d'un vecteur d'entrée à une classe audio ou visuelle respectivement, par analyse de la distribution des probabilités $\mathbf{P}^a[t] = [p^a_a[t], \ldots, p^a_{N_a}[t]]$ et $\mathbf{P}^v[t] = [p^v_1[t], \ldots, p^v_{N_v}[t]]$. La valeur de la décision au sein d'une modalité sera exprimée par l'index unique de la classe audio ou visuelle, respectivement, qui aura été considéré comme le vainqueur du processus de fusion. $d^a[t]$ et $d^v[t]$ seront donc définies dans l'intervalle $d^a \in \mathbb{N} = [1, N_a]$ pour l'audio et $d^v \in \mathbb{N} = [1, N_v]$ pour la vision.

Nous introduisons maintenant la définition de $d^c[t]$ comme la décision faite à partir de $d^a[t]$ et $d^v[t]$ grâce à laquelle la catégorie audiovisuelle de l'objet considéré sera déduite. A noter un fait important : $d^a[t]$ et $d^v[t]$ ne sont pas indépendants dans notre processus de fusion. Le fonctionnement de l'algorithme M-SOM, détaillé plus tard, lie les données audio aux données visuelles. Lorsque le système effectue l'analyse des données issues des experts d'identification la décision d^c ne consiste pas en la simple concaténation des décisions $d^a[t]$ et $d^v[t]$. La combinaison de la décision audio et visuelle sera détaillée une fois le M-SOM décrit.

Enfin, les définitions précédentes seront modulées par le fait que la catégorie a été inférée à partir d'une donnée incomplète ou qu'elle a été déduite d'une donnée complète. Ainsi, $d^a_{\text{MISS}}[n]$ et $d^v_{\text{MISS}}[n]$ signifiera que la décision a été prise à partir des données audio en absence de vision ou d'audio, respectivement. A l'opposé, $d^a_{\text{ALL}}[n]$ et $d^v_{\text{ALL}}[n]$ signifiera que la décision a été prise à partir d'une donnée audiovisuelle complète.

4.4.3 Catégorisation audiovisuelle

Le modèle HTM tente de faire émerger la notion d'*objet multimodal* par fusion des données issues des experts d'identification (cf. **Sec. 3.2.3**). Durant les expériences menées, et afin d'évaluer notre modèle, nous avons eu accès aux données réelles, c'est-à-dire aux catégories audiovisuelles auxquelles différentes sources présentes dans l'environnement appartenaient. A partir de cette connaissance, il a été possible de comparer les résultats de l'intégration multimodale du modèle HTM, et du module module MFI en particulier (le module DW n'effectuant pas d'intégration multimodale), avec les données réelles. Nous définissons ainsi les catégories suivantes :

Définition 7. *Catégorie réelle $\mathcal{C}_{real}$: Catégorie audiovisuelle d'une trame déterminée par la concaténation de la classification* réelle *des données audio et visuelles, i.e. sans erreurs des experts d'identification.*

Définition 8. *Catégorie estimée $\widehat{\mathcal{C}_{\text{MFI}}}$: Catégorie audiovisuelle d'une trame déterminée par l'analyse du module MFI à partir des données issues des experts d'identification audio et visuels. Ces données incluent ici d'éventuelles erreurs de classification.*

Définition 9. *Catégorie estimée $\widehat{\mathcal{C}_{\mathfrak{R}_n}}$: Catégorie audiovisuelle d'une trame déterminée par l'analyse du robot naïf à partir de la fusion brute des données issues des experts d'identification audio et visuels. Ces données incluent donc d'éventuelles erreurs de classification.*

Considérons, par exemple, $N_a = 3$ experts d'identification audio dédiés aux catégories $\mathcal{C}_1^a = \{\texttt{speech}\}$, $\mathcal{C}_2^a = \{\texttt{knock}\}$ et $\mathcal{C}_3^a = \{\texttt{alert}\}$. Similairement, considérons $N_v = 2$ experts d'identification visuels dédiés aux catégories $\mathcal{C}_1^v = \{door\}$ et $\mathcal{C}_2^v = \{female\}$. Ainsi, si une *femme* est en train de *parler* en face du robot, la catégorie audiovisuelle $\widehat{\mathcal{C}_{\text{MFI}}}$ déterminée par le module MFI est supposée être la même que la catégorie réelle $\mathcal{C}_{real} = \{\mathcal{C}_1^a, \mathcal{C}_2^v\}$. De plus, la catégorie $\widehat{\mathcal{C}_{\text{MFI}}}$ est supposée être la même que $\mathcal{C}_{real}$ même si (i) les experts d'identification audio et/ou visuels fournissent des données erronées ou (ii) si la donnée audio ou visuelle est manquante.

L'indice $\gamma[t]$ est calculé au temps t à partir de la comparaison entre la catégorie audiovisuelle $\widehat{\mathcal{C}}$ estimée par le module MFI ou le robot naïf d'un objet o_j, avec la catégorie réelle $\mathcal{C}_{real}$, selon :

$$\gamma(o_j)[t] = \begin{cases} 1 \text{ si } \widehat{\mathcal{C}}(o_j)[t] = \mathcal{C}_{real}(o_j)[t], \\ 0 \text{ sinon.} \end{cases} \tag{4.8}$$

Nous noterons γ_{MFI} l'indice calculé à partir de la fusion faite par le module MFI et $\gamma_{\mathfrak{R}_n}$ celle effectuée par le robot naïf. Cet indice nous permet de mesurer et d'observer l'évolution des performances de classification des systèmes étudiés. Nous avons également décidé de prendre en compte les résultats de classification passés en effectuant une moyenne glissante des taux de classification permettant ainsi d'apprécier la dynamique temporelle du système de fusion. Ainsi, pour chaque objet o_j, son taux de bonne classification $\mathbf{\Gamma}(o_j)[t]$, au temps t, sera donné selon :

$$\mathbf{\Gamma}(o_j)[t] = \mathbf{a} \times \sum_{k=t_i}^{t} \left(\widehat{\mathcal{C}}[k] = \mathcal{C}_{real}[k] \right) = \mathbf{a} \times \sum_{k=t_i}^{t} \gamma[k] \tag{4.9}$$

avec $\mathbf{a}$ un vecteur d'indices pouvant être exprimés de deux façons en fonction de la trame temporelle à partir de laquelle le calcul du taux de bonne classification démarre :

1. la première trame correspond à la première émission d'un son par la source $\mathcal{S}_j$:
$$\mathbf{a}' = 1/[1, ..., (t - t^1_{\mathcal{S}_j}) + 1] \tag{4.10}$$
où $t^1_{\mathcal{S}_j}$ représente cette première trame ;
2. la première trame correspond à la première catégorisation faite par le module MFI pour l'objet o_j :
$$\mathbf{a}'' = 1/[1, ..., (t - t_i) + 1] \tag{4.11}$$
où t_i représente cette première trame. Ainsi, si un objet apparait dans l'environnement au temps $t = 100$ mais qu'il est situé en-dehors du champ de vision du robot et que le modèle HTM n'est pas encore capable d'inférer sa catégorie audiovisuelle, cet objet n'aura aucun label audio, visuel ou audiovisuel. C'est à partir du moment où le modèle HTM assigne à l'objet une catégorie audiovisuelle pour la première fois que le taux $\mathbf{\Gamma}$ commencera à être calculé pour cet objet.

La première expression du vecteur $\mathbf{a}$, donnée par l'**Eq. 4.10**, permet d'observer la vitesse à laquelle le module MFI parvient à effectuer une bonne inférence et à explorer les différentes sources sonores présentes dans l'environnement. En effet, le temps entre le moment où une source se met à émettre un son et le moment où le module MFI effectue une fusion ou une inférence est ici pris en compte. La seconde expression du vecteur $\mathbf{a}$, donnée par l'**Eq. 4.11**, permet quant à elle de rendre compte directement des performances du module MFI en terme d'inférence et de convergence de l'apprentissage, en ne considérant que les résultats de classification à partir du moment où le module MFI a effectué sa première fusion / inférence de données.

Enfin, le taux de classification global est donné par la moyenne de tous les taux de classification par objet. Il est nécessaire de distinguer la catégorisation effectuée par le module MFI et celle effectuée par le robot naïf $\mathfrak{R}_n$. En effet, le fonctionnement du module MFI permet d'inférer la catégorie d'un objet sur la seule base des données audio. Ainsi, le module est capable de catégoriser un objet même lorsqu'il ne lui fait pas face. En revanche, le robot naïf lui ne peut pas effectuer cette inférence et ne pourra effectuer la fusion que des données auxquelles il a accès, c'est-à-dire celles de l'objet auquel il fait face. Cela se traduit par une expression du taux de bonne classification différente en fonction du système effectuant la fusion. Pour le module MFI, le taux de classification global est donné selon :

$$\bar{\mathbf{\Gamma}}^a_{\text{MFI}}[t] = \frac{1}{N^c_{obj}[t]} \sum_{j=1}^{N^c_{obj}[t]} \mathbf{\Gamma}(o_j)[t] \tag{4.12}$$

avec N^c_{obj} le nombre d'objets catégorisés par le module au temps t (ce nombre peut être inférieur ou égal au nombre d'objets présents), et l'exposant a pouvant être soit a' soit a'' (cf. **Eq. 4.10** & **Eq. 4.11** respectivement). Pour le robot naïf, en revanche, ce taux sera donné selon :

$$\bar{\mathbf{\Gamma}}_{\mathfrak{R}_n}[t] = \frac{1}{N_{obj}[t]} \sum_{j=1}^{N_{obj}[t]} \mathbf{\Gamma}(o_j)[t] \tag{4.13}$$

avec N_{obj} le nombre d'objets émettant au temps t.

Cependant, nous avons malgré tout souhaité comparer les performances du module MFI à la fusion des classifieurs effectuée par le robot naïf dans un cas où ce robot aurait tout le temps accès à toutes les données des classifieurs, que ce soient les données audio et visuelles. Il s'agit d'un cas irréaliste puisque le robot ne peut percevoir des données visuelles que s'il se trouve en face d'un objet. Ce taux de bonne classification sera exprimé de la même façon que l'**Eq. 4.13** mais sera noté $\bar{\mathbf{\Gamma}}'_{\mathfrak{R}_n}[t]$. Nous pourrons ici étudier l'intérêt de l'inférence de données manquantes par mesure vitesse à laquelle le module MFI rejoint puis dépasse les performances d'un système ayant accès à toutes les données, tout le temps. A noter ici que dans le cas unisource, nous aurons : $\bar{\mathbf{\Gamma}}'_{\mathfrak{R}_n}[t] = \bar{\mathbf{\Gamma}}_{\mathfrak{R}_n}[t]$. Le robot naïf aura donc deux modes de fonctionnement auxquels nous nous référerons tout au long de l'évaluation du modèle : un mode omniscient (irréaliste) dans lequel le robot a, à tout moment, un accès complet à toutes les données audiovisuelles — même les données visuelles sont situées en-dehors de son champ de vision — et un mode réaliste pour lequel il n'a accès qu'aux données complètes lorsqu'il fait face à la source émettant un son.

Enfin, nous aurons parfois à calculer le ratio entre les performances du module MFI et celles du robot naïf du point de vue des taux de classification, ratio que nous calculerons selon le rapport entre la moyenne des deux taux $\bar{\mathbf{\Gamma}}^a_{\text{MFI}}[t]$ (prenant en compte les deux expressions du vecteur $\mathbf{a}$, voir ci-dessus) et $\bar{\mathbf{\Gamma}}_{\mathfrak{R}_n}[t]$, le taux obtenu par le robot omniscient.

4.5 Conclusion du Chapitre

Ce chapitre est une introduction au modèle HTM permettant de poser les premières bases théoriques et conceptuelles sur lesquelles les trois chapitres suivants se basent. Les définitions d'un Robot, d'une Source, d'un Environnement, d'un Evénement, d'un Objet et de la Focalisation, et leurs notations respectives, ont été exposés, permettant de présenter les différents composants du problème auquel le modèle HTM tente de fournir une réponse : de quelle façon un robot peut-il réagir avec pertinence à la survenue d'événements émis par des sources présentes dans un environnement inconnu et aboutir à une représentation interne de cet environnement formalisée par les objets qui le composent. Cette notion d'Objet est donc une extension de la définition d'une Source en cela qu'elle intègre une dimension temporelle ainsi qu'un pendant sémantique qui sera détaillé au chapitre suivant : sa Congruence.

D'autre part, le simulateur développé pour tester le modèle a été décrit, simulateur nécessaire le temps d'avoir accès au robot entièrement fonctionnel et avec le logiciel Two!Ears totalement intégré. Le HtmTestBed consiste principalement en la simulation de la sortie des experts d'identification et de localisation grâce à la génération de vecteurs de probabilités similaires à ceux que les véritables experts génèrent. A

noter que la validation d'un modèle en environnement simulé est souvent plus aisée que sur une véritable plateforme robotique. C'est pourquoi nous avons pris soin de complexifier les données simulées, notamment en créant beaucoup plus d'experts de classification audio et visuels que le logiciel TWO!EARS ne possède, et en générant des vecteurs de probabilités bien plus bruités que les vecteurs réel. Cet environnement est également un outil de visualisation offrant la possibilité d'observer le robot au sein de l'environnement simulé, ainsi que ses mouvements de tête vers les différentes sources présentes. Différentes mesures des performances du modèle HTM sont également visibles, notamment le taux de bonne classification audiovisuelle ainsi que le nombre de mouvements de tête générés par le modèle et par le robot naïf. L'utilisation de cet environnement a été nécessaire durant la quasi totalité de cette thèse, le robot de l'ISIR n'ayant été pleinement fonctionnel qu'aux alentours du début du mois d'octobre 2016, période à laquelle le projet TWO!EARS se terminait. Cette période a d'ailleurs été consacrée à la rédaction de nombreux rapports finaux ainsi qu'à une série de tests préliminaires en vue de la préparation de la revue finale du projet, en janvier 2017, revue ayant eu lieu à Toulouse et donc sur le robot Jido. Malgré l'arrivée tardive d'Odi, l'intégration du modèle au sein de notre robot a pu être faite et les résultats obtenus seront présentés au **Chap. 7**. Mais même si le simulateur HtmTestBed a été un outil destiné à palier l'absence d'un robot fonctionnel, il nous a permis de tester le modèle dans des conditions beaucoup plus extrêmes que le robot doté du logiciel TWO!EARS n'aurait pu nous permettre, notamment dans des environnements massivement multisources (plus de 5 sources sonores).

Le chapitre suivant entre maintenant dans un des deux cœurs du modèle HTM : le module de Pondération Dynamique, permettant d'analyser un environnement inconnu composé de sources audiovisuelles et de générer une réponse comportementale bas-niveau en réaction à l'apparition de ces sources dans l'environnement. Nous souhaitons une nouvelle fois prévenir le lecteur qu'il sera probablement nécessaire de revenir à ce chapitre lors de la lecture de la suite du document.

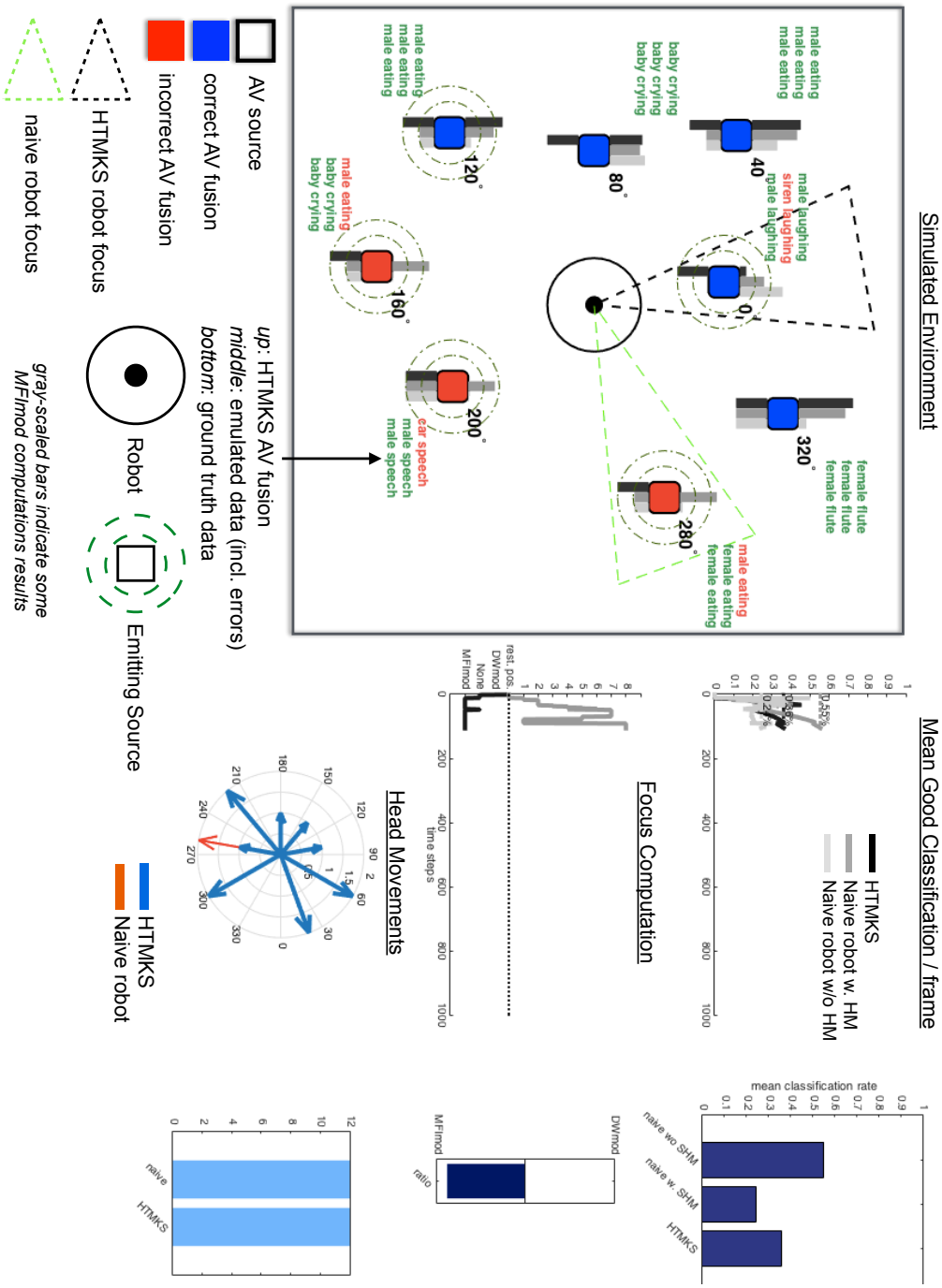


FIGURE 4.3 – ILLUSTRATION DU HTMTESTBED — Un environnement de simulation a été développé afin de tester les composants du modèle HTM séparément puis combinés. Ce simulateur a été nécessaire lorsque le système TWO!EARS était en développement et tant que le robot était indisponible. Une fois ces conditions réunies, le HtmTestBed a continué d'être intensivement utilisé pour la facilité avec laquelle il nous a permis de tester le modèle et de mesurer ses performances.

Chapitre 5

Module de Pondération Dynamique

Q. Comment définir un stimulus d'intérêt dans un environnement inconnu et sans règles données a priori *?*

Le module de Pondération Dynamique (module *Dynamic Weighting*, DW) a été le premier module implémenté dans le cadre du modèle HTM. Inspiré de phénomènes neurologiques et répondant aux besoins et contraintes du projet Two!Ears, le module DW tente d'adresser la question de l'*attention* du robot, concrétisée par les mouvements de sa tête, par une approche qui, bien qu'inscrite dans la continuité des travaux sur la saillance de signaux (cf. **Sec. 2.3.2**) est innovante : celle de la Congruence d'un événement étant donné l'environnement dans lequel celui-ci survient. Ce chapitre détaille l'ensemble du module DW.

La Sec. 5.1 décrit les concepts et considérations ayant motivé les choix faits pour la création du module DW.

La Sec. 5.2 introduit la notion de *Congruence* d'un objet audiovisuel dans un environnement en cours d'exploration.

La Sec. 5.3 détaille la formalisation du module : les définitions sur lesquelles il repose et les fonctions de pondération créées pour modéliser le phénomène de la *Congruence*.

La Sec. 5.4 présente le comportement attendu du module étant donné sa conceptualisation.

La Sec. 5.5 enfin, présente les résultats de l'exploration d'un environnement inconnu par un robot doté de mouvements de tête, en simulation, lorsque son comportement est soumis à l'analyse du module DW.

5.1 Motivations & Genèse

L'objectif du module *Dynamic Weighting* (DW) est de déclencher des mouvements de tête vers des objets détectés comme ayant une certaine *impor-*

tance pour la compréhension de l'environnement par le robot étant donné le contexte sémantique dans lequel le robot se trouve. Cet objectif fait appel aux questionnements du domaine du *filtrage attentionnel* (cf. **Sec. 2.3**), de la motivation à explorer un environnement (cf. **Sec. 2.1.3**) ainsi que de la façon dont les flux perceptifs sont analysés (cf. **Sec. 2.2.3**). La question de l'*importance* d'un événement a été l'objet de beaucoup d'attention durant la dernière décennie du fait, d'une part, de l'émergence de systèmes intelligents de plus en plus performants et, d'autre part, du nombre grandissant de robots interagissant avec des humains et nécessitant donc d'être capable de comprendre leur environnement de façon similaire à ceux-ci. En effet, doter un robot de la possibilté de savoir détecter, dans une scène audio, visuelle ou mixte, quel stimulus mérite son attention sans qu'aucune règle préalable sémantique lui ait été inculquée est une étape primordiale dans l'élaboration de ces robots « intelligents ». Le module DW tente ainsi d'apporter une réponse au problème de l'objet de l'attention d'un robot doté de mouvements de tête dans un environnement inconnu.

La question de l'attention d'un robot dans un contexte d'exploration d'environnements inconnus est primordiale et critique. Enormément de travaux cherchent à trouver une sorte de vérité dans les stimuli perçus par les capteurs du robot. En effet, nombre d'algorithmes tentent de détecter, dès l'émergence d'une perception sonore ou visuelle, les facteurs permettant de déduire des caractéristiques haut-niveau de l'événement perçu, comme la notion d'importance ou de saillance. Il serait même possible de définir comme axiome sur lequel tout le modèle HTM se base l'assertion suivante :

> *L'importance d'un objet n'est définie que par sa relation aux autre objets et non par une caractéristique intrinsèque présente dans les données que le robot perçoit de cet objet.*

Toutes les approches citées à la **Sec. 2.1.3** offrent de bons résultats et apportent des améliorations conséquentes aux algorithmes déjà existants. Cependant, beaucoup d'entre elles se basent sur des hypothèses fortes restreignant l'environnement exploré. L'ambition du module DW est de définir une approche beaucoup plus simple — sans être simpliste — ayant ainsi l'avantage de donner au robot une capacité d'adaptabilité très forte sans avoir de règles dictées préalablement par l'expérimentateur. Le module DW diffère ainsi des approches citées précédemment en plusieurs points :

1. **le contenu acoustique et visuel**, au sens des caractéristiques bas-niveau intrinsèque des stimuli, n'est pas pris en compte. Le module DW ne détermine pas l'importance d'un événement basé sur ses caractéristiques constitutives mais plutôt sur l'information plus haut niveau qu'il porte, i.e. sa catégorie audiovisuelle ;
2. le module DW est implémenté comme une boucle de rétroaction. Ainsi, l'expérience et la connaissance accumulées par le robot vont moduler sa capacité à faire émerger la notion d'*importance* d'un événement ;
3. la part **active** est primordiale puisqu'elle est, d'une part, ce qui va permettre au module DW d'apprendre l'environnement en cours d'exploration et, d'autre part, ce que le module va justement, à long terme, chercher à inhiber. En effet, le but du module DW est d'avoir une connaissance du

monde assez précise pour pouvoir ne générer des mouvements de tête que lorsque cela est nécessaire ;
4. il n'existe pas de « vérité » à apprendre sur l'environnement.

Afin d'effectuer cette pondération dynamique des événements apparaissant dans un environnement inconnu, le module DW va se baser sur les sorties des experts d'identification audio et visuels, ainsi que sur les estimations de la localisation des sources sonores (cf. **Sec. 3.2.3** pour la description de toutes les KS). A partir de ces sorties, le module va calculer la probabilité *a posteriori* que les objets audiovisuels ont d'apparaître à nouveau dans l'environnement en cours d'exploration. En fonction de la distribution de probabilité ainsi calculée, un ordre moteur sera déclenché vers l'objet d'intérêt. La section suivante est dédiée à l'introduction et à la formalisation de la notion de *Congruence*, concept liant l'apparition d'un événement, comme un objet audiovisuel, dans un environnement donné.

5.2 La notion de *Congruence*

Comme expliqué ci-dessus, le module DW a pour but de donner au robot la faculté d'assigner une mesure de l'*importance* à l'apparition d'objets audiovisuels lors de l'exploration de son environnement. Pour cela, nous introduisons la notion de *Congruence*. En algèbre, deux figures planes sont congrues [1] si elles ont des caractéristiques similaires, comme leur forme ou leur taille. Du point de bue biologique, la notion de congruence, et particulièrement son opposée, l'*incongruence*, est présente dans le système nerveux comme une réaction neuronale appelée *Mismatch Negativity* [167] (cf. **Sec. 2.3.1.4**). La capacité de prédiction des stimuli présentés par les aires sensiorielles provient d'une analyse des probabilités d'apparition calculées *a posteriori*. En effet, un stimulus peut-être interprété comme *déviant* dans une séquence donnée, mais être ensuite interprété comme *prédictible* dans une séquence différente au cours de laquelle ce stimulus est répété.

Nous proposons une extension de ce comportement neuronal à notre problématique. Si nous considérons que l'apparition d'objets constitue la séquence de stimuli prédictible au sein de laquelle peuvent se trouver des objets dont l'apparition serait imprédictible, l'analyse *a posteriori* de ces probabilités d'apparition devrait aboutir aux mêmes phénomènes d'habituation ou de sur-réaction. D'autre part, nous adapterons le phénomène de sur-réaction à un stimulus imprédictible formalisé par l'onde N100 dans les aires sensorielles, aux mouvements de tête du robot. Ainsi, l'apparition prédictible d'un objet dans un environnement inconnu ne devrait susciter aucune réaction motrice de la part du robot, tandis qu'une apparition imprédictible devrait provoquer un mouvement de tête vers cet objet. Beaucoup de modèles computationnels d'exploration supervisée par des motivations attentionnelles placent au centre de leur analyse de la scène perceptive la *Saillance* (définie à la **Sec. 2.3.1.2**) des événements.

En quoi la *Congruence* diffère-t-elle de la *Saillance* ?

La distinction effectuée ici tient de la définition de ce qu'est un événement susceptible

1. également *congruentes*.

de réquérir une modification de l'état attentionnel d'un robot. Selon une définition basée sur la saillance, ce qui doit requérir l'attention du robot doit être le signal qui porte en ses caractéristiques bas-niveau un marqueur de différence par rapport à son environnement *temporel* à court terme (puisqu'il n'y a pas de phénomène d'habituation à moyen ou long terme). Selon une définition basée sur la Congruence, ce qui doit requérir l'attention du robot doit être le signal qui porte en son *sens*, un marqueur de différence par rapport à son environnement *sémantique*. C'est cet aspect sémantique, basé sur la notion d'information, qui est au centre du module DW. De plus, l'aspect temporel est conservé puisque le « sens » d'un événement audiovisuel dépendra de son environnement, donc des événements audiovisuels déjà perçus. Nous pouvons alors définir la Congruence de l'apparition d'un Evénement au sein d'un Environnement comme suit :

Définition 10. *Congruence :*

1. *caractéristiques de deux événements entraînant une perception similaire ;*
2. *propriété liant un événement perçu et son environnement (i.e. les événements perçus précédemment).*

La Congruence d'un Evénement peut donc être vue comme un *état* particulier de celui-ci, dépendant des Evénements précédents. La dimension temporelle de la Congruence est majeure car constitutive. Ainsi, une des propriétés majeures de la Congruence concerne sa convergence temporelle :

Propriété 2. *Convergence : La valeur de la congruence d'un événement ne converge pas temporellement. Etant fonction de l'apparition d'autres événements, elle est toujours susceptible d'être modifiée et de converger vers son opposé (l'Incongruence).*

Ainsi, un objet congru peut devenir incongru, et inversement. La **Propr. 2** est extrêmement importante car elle reflète le caractère adaptatif de la Congruence telle que nous l'utilisons. Avec la **Déf. 10** nous tirons que la Congruence n'est pas une caractéristique figée et absolue d'un objet mais est directement dépendante de (i) l'environnement dans lequel il se situe et (ii) des autres objets présents à un temps t donné de l'exploration de l'environnement. Cette définition nous éloigne de la notion de saillance telle qu'elle est souvent utilisée dans les travaux de recherche (comme ceux décrits à la **Sec. 2.3.2**) car elle ne considère pas qu'un événement acoustique ou visuel soit saillant par essence. La notion de Congruence que nous proposons ne doit pas être comprise comme une opposition à la saillance ou une réfutation de la pertinence de l'analyse de scènes audio ou visuelles par le prisme de la saillance. La Congruence est *une forme de saillance*. Seul le contenu environnant l'entité considérée change. Par exemple, dans une scène visuelle, si un ensemble de balles rouges sont lancées régulièrement dans une pièce de couleur pâle à des endroits divers, toutes ces balles rouges seront détectées comme des entités saillantes et pourront, dans le cas des modèles cités à la **Sec. 2.3.2**, déclencher une réaction motrice dans leur direction, comme des mouvements oculaires par exemple. Dans le cas de la Congruence, les toutes premières apparitions de ces objets aboutiraient à une représentation de ceux-ci comme *saillants* mais très rapidement, étant donné le caractère répétitif de

ces événements, l'apparition de nouvelles balles rouges serait considérée comme un événement *non saillant*. Sur la base de cette analyse des événements audiovisuels que le robot va percevoir, un mouvement de tête va éventuellement être généré, d'une façon similaire des saccades oculaires générés par les modèles de saillance visuelle décrits précédemment.

La section suivante présente la formalisation de la notion de Congruence ainsi que celle de l'ordre moteur consécutif à cette analyse.

5.3 Formalisation

Le module DW est en charge de l'apprentissage des probabilités *a posteriori* des événements audiovisuels perçus par le robot. A chaque nouvel événement détecté par le robot, le module DW calcule sa Congruence étant donné l'état de l'exploration de l'environnement exprimé par les objets audiovisuels qui le composent et qui ont été détectés par le robot. Cette analyse de la Congruence de l'apparition d'un objet sera formalisée par une fonction de pondération permettant d'assigner des poids aux différents objets détectés. A partir de ces poids, un ordre moteur pourra éventuellement être généré dans la direction de l'objet d'intérêt. Ce mouvement permettra de focaliser l'attention du robot sur cet objet en mobilisant ses capteurs visuels sur lui. La formalisation du module DW consiste en (i) la description du processus de pondération et (ii) celle de la génération des mouvements de tête.

5.3.1 Pondération

Pour décider si un objet o_j apparaissant dans un Environnement requiert l'attention du robot, un poids va lui être assigné, poids étant le résultat de l'analyse de la Congruence de cet objet à son environnement. Nous définissons ainsi la congruence d'un objet comme :

Définition 11. *Congruence d'un objet : o_j est défini comme* congru *si suffisamment d'autres objets appartenant à la même catégorie $c^{(i)}(a_j, v_j)$ ont déjà été détectés précédemment par le robot.*

---

et l'incongruence d'un objet comme :

Définition 12. *Incongruence d'un objet : o_j est défini comme* incongru *si d'autres objets appartenant à la même catégorie $c^{(i)}(a_j, v_j)$ n'ont jamais, ou très peu, été détectés précédemment par le robot.*

Une fonction de pondération de l'apparition d'un objet dans un environnement $f_\omega(p) \in \mathbb{Z} = [-1, 1]$ a ainsi été conçue de telle sorte qu'un objet incongru aura un

poids $w(o_j) = 1$, tandis qu'à l'inverse, un objet congru aura un poids $w(o_j) = -1$. En fonction de la valeur des poids de tous les objets présents dans l'environnement, un mouvement de tête vers les objets les plus incongrus (ceux ayant le poids le plus élevé) sera déclenché.

Sur la base de ces définitions, et en se plaçant à un temps discret t de l'exploration, nous définissons la pseudo-probabilité $p(c^{(i)}(a_j, v_j))$ comme :

$$p\left(c^{(i)}(a_j, v_j)\right) = \frac{|c^{(i)}(aj, v_j)|}{N_i} \tag{5.1}$$

où $|c^{(i)}(a_j, v_j)|$ représente le nombre d'objets appartenant à la catégorie $c^{(i)}(a_j, v_j)$ et N_i le nombre total d'objets. De plus :

$$\sum_{n=1}^{|\mathcal{C}^{(i)}|} p(c^{(i)}(a_n, v_n)) = 1 \tag{5.2}$$

avec $|\mathcal{C}^{(i)}|$ le nombre de catégories détectées par le robot au temps t. La pseudo-probabilité $p(c^{(i)}(a_j, v_j))$ peut être considérée comme la probabilité qu'un objet o_j appartienne à la catégorie $c^{(i)}(a_j, v_j)$.

La fonction de pondération $f_\omega(p) : \mathbb{R}^+ \to \mathbb{Z} = [-1, 1]$, illustrée à la **Fig. 5.1**, associe un poids $w(o_j)$ à une probabilité selon :

$$f_\omega(p) = \begin{cases} 1 & \text{si } p <= K_i, \\ -1 & \text{sinon.} \end{cases} \tag{5.3}$$

où K_i représente l'équiprobabilité d'apparition des catégories audiovisuelles :

$$K_i = \frac{1}{|\mathcal{C}^{(i)}|} \tag{5.4}$$

Ce seuil a été choisi afin de respecter le principe d'absence de règles données *a priori* au robot. Ainsi, nous évitons au maximum d'introduire un biais dans la fréquence attendue d'apparition des catégories audiovisuelles. L'équiprobabilité est un critère minimisant ce biais puisque toutes les catégories ont une probabilité égale d'apparaître dans un environnement donné. La valeur attribuée par la fonction de pondération sera ensuite affectée à l'objet o_j considéré selon :

$$w(o_j) = f_\omega(p(c^{(i)}(a_j, v_j))) \tag{5.5}$$

L'**Eq. 5.3** montre la relation entre un poids $w(o_j)$ élevé et une probabilité faible d'apparition de la catégorie à laquelle l'objet appartient $p(c^{(i)}(a_j, v_j))$. Ainsi, si un objet o_j apparaît dans l'environnement en cours d'exploration et qu'il appartient à une catégorie peu détectée jusqu'à présent, il sera défini comme *incongru* et un mouvement de tête sera alors déclenché dans sa direction. De cette équation, nous proposons une nouvelle propriété de la Congruence :

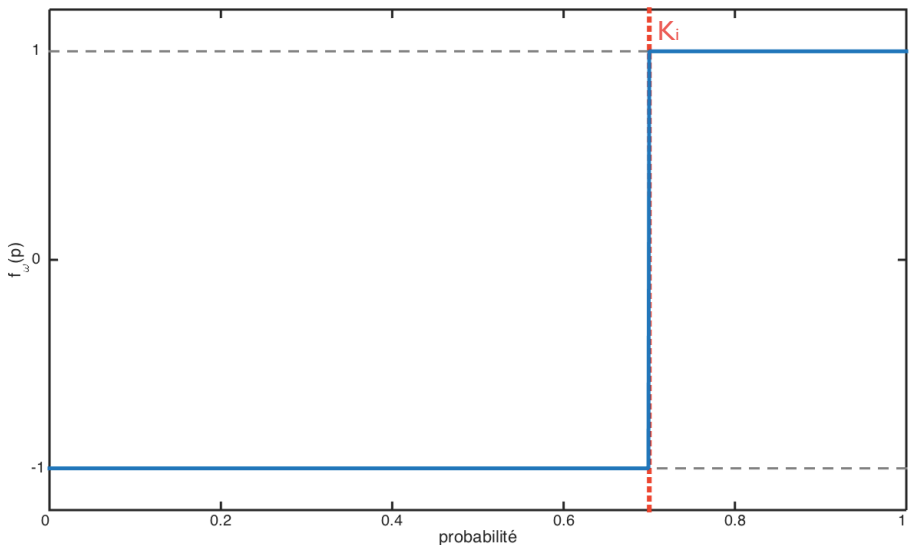


FIGURE 5.1 – FONCTION DE PONDÉRATION — Illustration du comportement de $f_\omega(p)$ associant à la probabilité $p(c^{(i)}(a_j, v_j)$ qu'un objet a d'appartenir à une catégorie audiovisuelle $c^{(i)}(a_j, v_j)$ à un poids $w(o_j) \in \mathbb{Z} = [-1, 1]$ (cf. **Eq. 5.3**)

Propriété 3. *Objet Congru : Un objet a un poids $w(o_j) = -1$ et est défini comme* congru *si la probabilité $p\left(c^{(i)}(a_j, v_j)\right)$ est supérieure à l'équiprobabilité.*

et de son opposé :

Propriété 4. *Objet Incongru : Un objet a un poids $w(o_j) = 1$ et est défini comme* incongru *si la probabilité $p\left(c^{(i)}(a_j, v_j)\right)$ est inférieure à l'équiprobabilité.*

La fonction de pondération $f_\omega(p)$ présentée par l'**Eq. 5.3** est ainsi similaire à une fonction carrée ayant deux valeurs possibles : -1 ou 1. Ce comportement est cependant trop binaire, particulièrement lorsque plusieurs objets sont présents simultanément. De plus, des erreurs de classification des experts d'identification peuvent survenir ponctuellement entraînant une erreur dans les catégories sur lesquelles le module DW se base et ainsi fausser le calcul de la Congruence. Cela engendrerait ainsi des mouvements de tête (i) erronés et donc non pertinents, et (ii) plus nombreux du fait de l'instabilité des données en entrée. Pour prendre en compte ces erreurs, la fonction de pondération $f_\omega(p)$ a été modifiée afin d'y introduire une intégration temporelle, procédé courant lorsqu'il est nécessaire de consolider les données sur lesquelles un système se base, permettant de rendre plus robuste le calcul de Congruence. Ainsi, nous avons transformé la fonction $f_\omega(p)$ en deux fonctions sigmoïdes séparées, au comportement symétrique, notées f_ω° et $f_\omega^\bullet$. Chacune de ces fonctions est dédiée à la pondération d'un objet congru ou incongru, respectivement. Elles convergent vers leur asymptote, inchangée, en $n = 5$ pas de temps. Le choix de cette valeur a été initialement motivé par deux considérations :

1. la dynamique temporelle de l'onde N100, apparaissant au bout d'environ 100 ms ;
2. la taille traditionnellement utilisée dans l'analyse des signaux audio (et utilisable pour l'analyse des signaux visuels également) d'une valeur de 20 ms.

Prenant en compte ces deux points, nous obtenons une valeur de $n = 100/20 = 5$ pas de temps pour converger vers la valeur maximale de ces deux fonctions. Cependant, ces deux considérations ont été faites au tout début du projet TWO!EARS et avant que la fenêtre d'intégration temporelle des différents experts d'analyse bas-niveau des signaux n'ait été choisie. Prenant en compte la dynamique temporelle du BLACKBOARD tel qu'il existe dans sa version finale, chaque pas de temps d'intégration est de 500 ms. Nous avons cependant gardé la valeur de $n = 5$ pas de temps pour l'intégration temporelle.

Les deux nouvelles fonctions de pondération sont désormais définies dans l'intervalle $f_\omega^{(\bullet/\circ)} : \mathbb{R}^+ \to \mathbb{R} = [-1, 1]$. Un autre changement a également été nécessaire : ces fonctions ne sont plus directement dépendantes de la probabilité $p\left(c^{(i)}(a_j, v_j)\right)$ mais d'un index temporel n défini comme *le nombre de trames durant lequel l'objet a été classé comme congru ou incongru.* Ainsi, ces deux fonctions ont pour expressions :

$$f_\omega^\circ(n) = (1/1 + 0.01\, e^{2n}) - 1 \tag{5.6}$$

pour la pondération d'un événement congru, et :

$$f_\omega^\bullet(n) = 1/(1 + 100\, e^{-2n}) \tag{5.7}$$

pour la pondération d'un événement incongru. La **Fig. 5.2** illustre le comportement de ces deux fonctions. Par exemple, un objet ayant été détecté comme **congru** pendant $n = 3$ trames consécutives, la valeur de son poids $w(o_j)$ au temps t sera de $f_\omega^\circ(3) \approx -0.8014$. A l'inverse, s'il est détecté comme **incongru** pendant $n = 5$ trames consécutives, la valeur de son poids $w(o_j)$ au temps t sera de $f_\omega^\bullet(5) \approx 0.9955$.

Les conditions d'utilisation de l'utilisation des fonctions $f_\omega^\bullet$ et f_ω° sont similaires à l'**Eq. 5.3** :

$$w(o_j)[t] = \begin{cases} f_\omega^\bullet(n) & \text{si } p\left(c^{(i)}(a_j, v_j)\right) \leq K_i, \\ f_\omega^\circ(n) & \text{sinon.} \end{cases} \tag{5.8}$$

Enfin, nous notons $\mathbf{W}^{(i)}[t]$ comme le vecteur contenant les valeurs de congruence de l'ensemble des catégories audiovisuelles $\mathcal{C}^{(i)}$ de l'environnement $e^{(i)}$ au temps t, selon :

$$\mathbf{W}^{(i)}[t] = [w(c_1^{(i)})[t], \ldots, w(c_K^{(i)})[t]] \tag{5.9}$$

Cette notation nous servira lors de la formalisation de la transmission des connaissances (détaillée plus tard) en cela que $\mathbf{W}^{(i)}[t]$ représente les *règles de Congruence* que le module DW va appliquer dans l'environnement $e^{(i)}$, règles qui seront susceptibles d'être transmises à un autre environnement.

Pour finir, il est important de préciser le comportement du module DW durant les toutes premières apparitions d'événements audiovisuels. Etant donnée l'**Eq. 5.8**, le poids associé à un objet perçu sera positif si la probabilité *a posteriori* d'observer

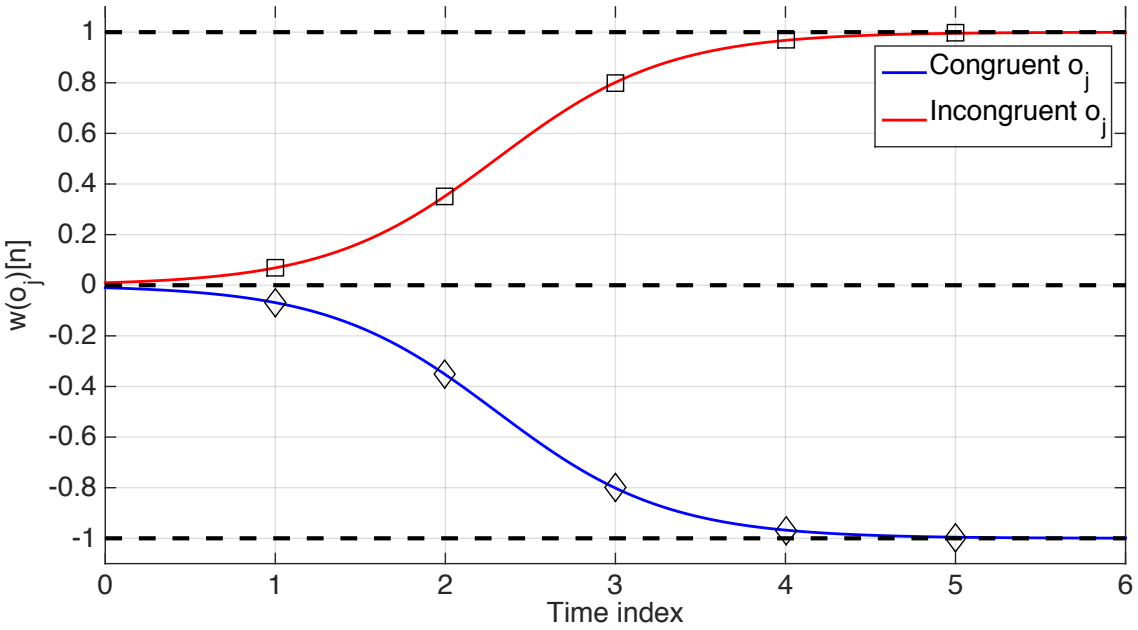


FIGURE 5.2 – FONCTIONS DE PONDÉRATION — (*bleu*) fonction positive f°_{ω}, (*rouge*) fonction négative $f^{\bullet}_{\omega}$. En fonction du caractère *congru* ou *incongru* de l'événement perçu, le poids qui lui sera assigné sera calculé selon une de ces deux fonctions.

l'apparition de cet objet est *inférieure ou égale* au critère K_i. Cela implique que lorsque le premier objet survient dans l'environnement, il sera catégorisé comme *incongru* puisque sa probabilité sera égale à K_i. Par extension, si tous les objets survenant dans l'environnement appartiennent tous à cette même catégorie, le module DW les catégorisera tous comme incongrus. Il aurait tout à fait été possible de définir une inégalité stricte, ce qui aurait eu pour conséquence de ne pas générer de mouvement de tête vers les tous premiers objets apparaissant. Cependant, puisque nous nous situons dans un contexte d'exploration de l'environnement, le choix a été fait de favoriser, dans cette situation, la génération d'un ordre moteur dans la direction de l'objet.

Le calcul des poids et leur assignation aux objets perçus par le robot constitue la base sur laquelle un ordre moteur est généré ou non, objet de la section suivante.

5.3.2 Ordre moteur

En fonction des valeurs des poids de tous les objets présents dans l'environnement et perçus par le robot, à un temps t, une décision sur un éventuel mouvement de tête doit être prise. Soit $\hat{\theta}^a_j$ l'estimation de la position angulaire de l'objet o_j obtenue par la localisation audio (exprimée par l'exposant ^{a}). L'angle de rotation de la tête requis par le module DW sera ainsi déterminé selon la propriété suivante :

Propriété 5. *Un ordre moteur est généré pour tourner la tête du robot d'un angle $\theta_m[t]$ si et seulement si un des événements a été détecté comme incongru.*

Or plusieurs événements peuvent être considérés en même temps comme incongrus au temps t. Il est donc nécessaire d'en sélectionner un parmi ceux-ci. Pour cela,

nous nous sommes inspirés de la boucle ganglions de la base — thalamus — cortex, introduite à la **Sec. 2.1.1** et jouant un rôle dans le mécanisme de sélection de l'action motrice. Notamment, lorsqu'une contradiction entre deux ordres moteurs existe, par exemple aller à gauche et à droite en même temps, cette boucle neurale permet de résoudre le conflit. Nous nous sommes plus particulièrement inspirés du modèle GPR [52, 53], introduit à la **Sec. 2.1.2.2** selon lequel une action motrice est formalisée par un canal d'information dont l'activité est plus ou moins inhibée, permettant ainsi d'effectuer un filtrage des actions motrices possibles et d'en sélectionner une.

Nous représentons chaque événement comme un canal d'information similaire à ceux des ganglions de la base, canal possédant une activité propre. Soit un événement ψ_j appartenant à la classe $c^{(i)}(a_j, v_j)$ et assigné à un objet o_j, nous définissons l'activité de chaque canal au temps t selon :

$$\tau_{\text{DW}}(\psi_j)[t] = -\frac{p\left(c^{(i)}(a_j, v_j)\right)}{K_i} \tag{5.10}$$

avec K_i étant l'équiprobabilité, définie à l'**Eq. 5.4**. Ainsi, plus un objet a un poids élevé, plus l'activité du canal correspondant sera faible. Nous notons τ_{DW} l'ensemble des activités des J événements perçus au temps t donné par la concaténation de tous les $\tau_{\text{DW}}(\psi_j)[t]$.

Définissons maintenant $\mathbf{\Theta}_{\text{DW}} \in \mathbb{R}^+ = [0, 359]$ comme l'ensemble des ordres moteurs possibles vers les J événements perçus par le robot au temps t, tel que :

$$\mathbf{\Theta}_{\text{DW}}[t] = [\theta(\psi_1)[t], \ldots, \theta(\psi_j)[t], \ldots, \theta(\psi_J)[t]] \tag{5.11}$$

où chacun des angles est donné par l'expert de localisation audio. Soit maintenant :

$$\tau_{min} = \arg\min_J(\tau_{\text{DW}}) \tag{5.12}$$

l'activité du canal la plus faible. Par suite, l'ordre moteur d'angle $\theta_{\text{DW}}[t]$ sera donc déterminé selon :

$$\theta_{\text{DW}}[t] = \mathbf{\Theta}_{\text{DW}}(\psi_{min}) \tag{5.13}$$

où ψ_{min} est l'événement ayant l'activité la plus faible.

Enfin, dans le cas où deux canaux auraient la même activité, c'est-à-dire deux objets ayant le même poids, l'activité sera alors modulée par un facteur temporel. Soit $t_i(o_j)$ le temps auquel un objet se met à émettre, nous définissons alors $\Delta_t(o_j)$:

$$\Delta_t(o_j) = t - t_i(o_j) \tag{5.14}$$

Ainsi, et seulement dans le cas où deux objets ont le même poids :

$$\tau_{\text{DW}}(\psi_j)[t] = -\frac{p\left(c^{(i)}(a_j, v_j)\right)}{K_i[t]} \times \frac{1}{\Delta_t(o_j)} \tag{5.15}$$

L'inclusion de $\Delta_t(o_j)$ dans le calcul de l'activité arrivera souvent puisque les poids des objets auront tendance à converger assez rapidement vers leurs valeurs extrêmes. L'**Eq. 5.15** permet d'inclure la motivation par la Nouveauté — forme de Curiosité — en cela que lorsque plusieurs événements attirent l'attention du robot, celui-ci choisira le plus « nouveau », i.e. celui qui est apparu le plus récemment.

Cette formalisation de la génération des ordres moteurs par le module DW, inspirée du modèle GPR, permettra ainsi de sélectionner la commande motrice « gagnante » lorsque plusieurs sont possibles.

5.3.3 Transmission des connaissances

Le module *Dynamic Weighting* est un module conférant au robot la capacité de réagir très rapidement à l'apparition d'événements audiovisuels dans un environnement inconnu. Mais le module DW ne consiste pas qu'en un module attentionnel bas-niveau : il inclut également une forme de mémoire. Notre définition d'un Environnement (cf. **Déf. 5**) stipule qu'il est défini par les objets audiovisuels qui le composent. Nous ajoutons à cette définition une précision importante : par « ensemble des objets audiovisuels », nous comprenons autant la qualité que la *distribution* des catégories auxquelles ces objets appartiennent. Au sein du module DW, cette distribution est exprimée par la probabilité *a posteriori* d'un objet appartenant à une catégorie d'apparaître dans un environnement. Cette probabilité est fonction de toutes les autres catégories audiovisuelles ayant été observées dans un environnement donné. En conséquence, l'ensemble des règles de Congruence apprises lors de l'exploration d'un environnement font également partie de la définition de cet environnement. Ces règles sont formalisées par le vecteur $\mathbf{W}^{(i)}[t]$ défini par l'**Eq. 5.9**, contenant les valeurs des poids assignés à chaque catégorie au temps t.

Ainsi, à chaque fois que le robot explore un nouvel environnement inconnu, il créera un nouvel ensemble de règles de Congruence dédiées à cet environnement. Cependant, l'intérêt d'avoir une mémoire — ou un ensemble de connaissances accessible à tout moment — est de pouvoir l'utiliser dans des situations nouvelles afin d'accélérer l'analyse de ces situations. Nous avons ainsi permis au module DW d'appliquer les connaissances aquises dans des environnements déjà explorés à tout nouvel environnement en cours d'exploration. De la même façon que l'ensemble du module DW (et globalement, l'ensemble du modèle HTM), une des plus fortes contraintes est de pouvoir effectuer une analyse nécessitant le moins de données et pouvant générer une réaction le plus rapidement possible. Suivant cette contrainte, la formalisation que nous proposons ne doit pas requérir d'avoir exploré un nombre conséquent d'environnements avant de pouvoir appliquer une quelconque connaissance. A l'extrême, à partir du deuxième nouvel environnement inconnu, le robot doit pouvoir tenter d'appliquer les connaissances qu'il a déjà aquises.

Afin de donner au robot la capacité de transmettre ses connaissances d'un environnement à l'autre, le module DW va, à chaque fois qu'un nouvel objet audiovisuel appartenant à une catégorie audiovisuelle donnée apparait dans un environnement en cours d'exploration, étudier la ressemblance entre la distribution des catégories audiovisuelles de cet environnement et la distribution des catégories des environnements passés. Cette ressemblance est formalisée par une règle d'inclusion entre

les différents ensembles de catégories $\mathcal{C}^{(i)}$ [2] des N_e environnements explorés jusqu'à présent par le robot, selon la définition suivante :

Définition 13. *Inclusion Stricte d'un Ensemble de Catégories : Les règles de Congruence* $\mathbf{W}^{(j)}$ *de l'environnement* $e^{(j)}$ *peuvent être appliquées à l'environnement* $e^{(i)}$ *en cours d'exploration si et seulement si* $\mathcal{C}^{(i)} \subsetneqq \mathcal{C}^{(j)}$.

$A \subsetneqq B$ dénote la relation d'inclusion $A \subseteq B$ ainsi que la distinction des deux ensembles $A \neq B$. S'en suit sa corollaire :

Corollaire 1. Transmission des Règles de Congruence : *Si* $\mathcal{C}^{(i)} \subsetneqq \mathcal{C}^{(j)}$, *alors* $\mathbf{W}^{(i)} = \mathbf{W}^{(j)}$.

Dans le cas où l'inclusion stricte est multiple, c'est-à-dire si l'ensemble $\mathcal{C}^{(i)}$ est inclus dans plusieurs autres ensembles $\mathcal{C}$, l'environnement sélectionné sera celui qui a été exploré le plus récemment, environnement ayant déjà éventuellement subi une transmission de connaissances d'un autre environnement.

5.3.4 Discussion

Nous avons, dans cette section, posé les bases du module module DW : pondération d'un objet à partir du calcul de sa Congruence, ordre moteur généré à la suite de cette pondération des objets présents et transmission des connaissances d'un environnement à l'autre. L'ensemble de la formalisation du module DW, bien qu'assez simple, constitue un ensemble théorique permettant de donner à un robot évoluant dans un environnement inconnu et sans règles comportementales données *a priori*, la possibilité d'analyser l'environnement dans lequel il se situe et d'y réagir. Cet environnement, nous l'avons défini par l'ensemble des objets qui le composent et, par extension, par la distibution des probabilités *a posteriori* d'apparition des catégories audiovisuelles auxquelles ces objets appartiennent. L'inspiration des phénomènes neuronaux ou, plus globalement, cérébraux, et détaillés au **Chap. 2**, nous ont permis de modéliser certains de ces comportements de façon simple et sans verser dans le biomimétisme. En effet, une approche biomimétique consisterait par exemple en l'implémentation des neurones du colliculus supérieur recevant des afférences des récepteurs audio et visuels et générant des commandes motrices vers le cou et les yeux, notamment. Nous avons préféré rester dans la sphère « bio-inspirée » en cela que nous pensons que les mécanismes éminemment complexes du cerveau peuvent, dans un premier temps, être modélisés grâce à des formalisations et des concepts plus simples, et la Congruence d'un objet dans un environnement basé sur le calcul de probabilités *a posteriori* en fait partie.

Le module DW étant maintenant formalisé, la section suivante décrit le comportement attendu du module.

---

2. Nous rappelons ici que $\mathcal{C}^{(i)} = \{c^{(i)}(a, v)\}$ représente l'ensemble des catégories du i-ème environnement tel que $\mathcal{C}^{(i)} = \{c^{(i)}(a, v)\}$, cf. **Sec. 4.1.2**.

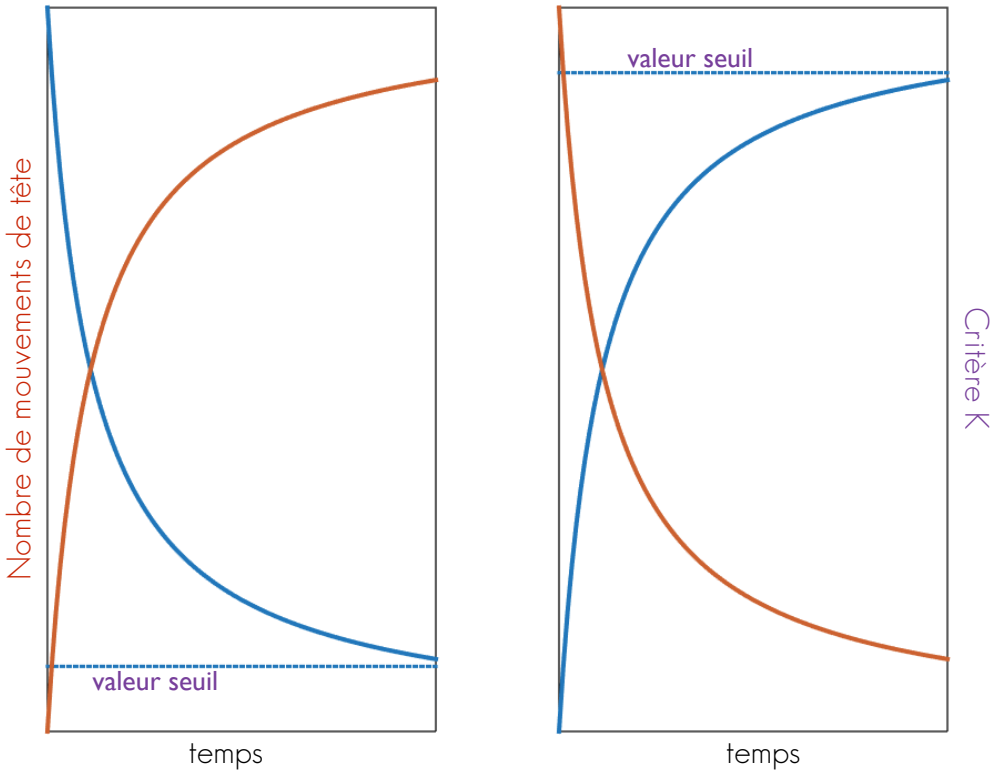


FIGURE 5.3 – COMPORTEMENT ATTENDU DU MODULE DW — (*gauche*) Cas extrême où chaque nouvel objet apparaissant dans l'environnement appartiendrait à une catégorie audiovisuelle nouvelle. Le critère K_i, directement fonction de l'équiprobabilité d'apparition des catégories (cf. **Eq. 5.4**), convergerait ainsi vers une valeur de plus en plus basse, entrainant des mouvements de tête à chaque nouvelle apparition d'un objet ; (*droite*) cas extrême où la quasi totalité des objets apparaissant dans l'environnement appartiendrait à la même catégorie. Le critère K_i convergerait alors vers une valeur de plus en plus haute, inhibant ainsi tous les mouvements de tête.

5.4 Comportement attendu

A PARTIR de la description que nous venons de faire du module DW, il est possible d'émettre une hypothèse sur le comportement attendu du module en fonction des événements apparaissant dans l'environnement en cours d'exploration. La **Fig. 5.3** illustre ce comportement dans deux cas extrêmes et idéalisés :

(*gauche*) chaque nouvel événement apparaissant dans l'environnement appartient à une nouvelle catégorie audiovisuelle $\mathcal{C}$. Le critère $K_i = 1/|\mathcal{C}^{(i)}|$ (cf. **Eq. 5.4**) convergerait vers une valeur de plus en plus basse, provoquant un mouvement de tête à chaque apparition d'un événement. Ce comportement est le même que celui du robot naïf $\mathfrak{R}^{\mathrm{n}}$ (cf. **Déf. 1**).

(*droite*) quasiment tous les événements apparaissant dans l'environnement appartiennent à la même catégorie $\mathcal{C}$. Le critère K_i converge alors vers une valeur de plus en plus haute, entrainant une inhibition quasi complète des mouvements de tête par caractérisation de chaque événement comme *congru*.

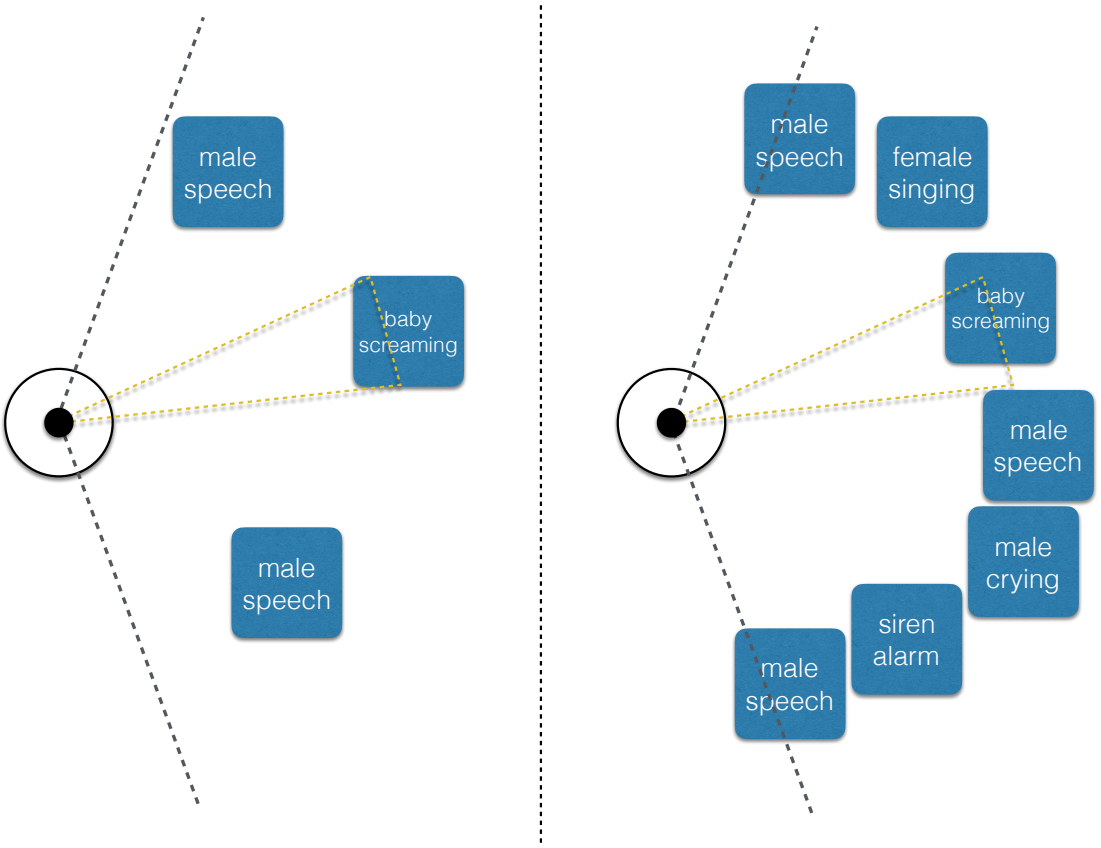


FIGURE 5.4 – DW, SCHÉMA SIMPLIFIÉ D'ENVIRONNEMENTS SIMULÉS — Exemple de deux scénarios simulés grâce au HtmTestBed et intégrant les contraintes de position des sources. (*traits noirs pointillés*) champ de vision simulé (celui est démesurément grand, volontairement) ; (*triangles verts pointillés*) direction de la tête.

5.5 Resultats

LE MODULE module DW a été évalué par sa capacité à générer des mouvements de tête vers des sources d'intérêt sur la base de la formalisation de la notion de Congruence définie ci-dessus. Le robot simulé ici a été comparé au robot naïf $\mathfrak{R}_n$ (cf. **Déf. 1**) tournant la tête à chaque fois qu'un événement audiovisuel apparait dans l'environnement.

Afin d'évaluer le module DW nous avons généré de nombreux scénarios grâce au HtmTestBed (cf. **Sec. 4.3**). Le fait d'avoir développé ce simulateur nous permet de tester le modèle dans des conditions extrêmes, conditions que le système TWO!EARS (ni aucun autre du même type, à ce jour) ne saurait gérer, notamment du fait de la présence de nombreuses sources émettant simultanément. Chaque scénario consiste en la simulation d'un environnement peuplé de $n_{\mathcal{S}}$ sources audiovisuelles $\mathcal{S}$ auxquelles sont associées des catégories $\mathcal{C}^{(a,v)}$. Les données récupérées puis analysées par le module DW proviennent des KS simulées par le HtmTestBed. Un scénario est défini par le nombre de sources audiovisuelles présentes, le nombre maximum autorisé de sources émettant du son de façon simultanée, la durée de la simulation et le type de catégories audiovisuelles présentes. Une catégorie audiovisuelle peut être associée

à une ou plusieurs sources $\mathcal{S}$. Il s'agira de plusieurs objets distincts mais dont les experts d'identification audio et visuels auront assigné les mêmes catégories : ainsi, deux (ou plusieurs) sources peuvent être étiquettées *male speech* par exemple.

D'autre part, nous avons modifié volontairement l'environnement dans lequel le robot simulé se situe : nous nous plaçons dans un cas où le robot a un champ de vision de 180°, lui permettant d'avoir tout le temps accès aux données visuelles (cf. **Fig. 5.4**). Ainsi, un mouvement de tête requis par le module DW vers une source correspondra au mouvement nécessaire pour que la source d'intérêt soit à l'angle 0°. Cela ne change absolument rien au fonctionnement du DW mais nous verrons à la fin de ce chapitre que le non-respect de cette condition, hautement probable dans des scénarios réalistes, a été une des limites du module DW et une des motivations principales du développement du module MFI, décrit au chapitre suivant.

5.5.1 Comportement global

Pour observer la dynamique globale du module DW et de l'exploration de la scène audiovisuelle modulée par la Congruence assignée aux objets perçus, nous avons créés trois scénarios, détaillés au **Tab. 5.1**.

Conditions de test 5.5.1 [3]				
n°	$n_{\mathcal{S}}$	n_{sim}^{max}	T	Catégories présentes [4]
1	3	1	500	1, 28
2	6	4	500	1, 9, 28
3	10	7	500	1, 3, 9, 28, 41, 46

TABLE 5.1 – Caractéristiques des trois scénarios générés pour observer le comportement global du module DW. Ces trois scénarios sont d'une complexité croissante.

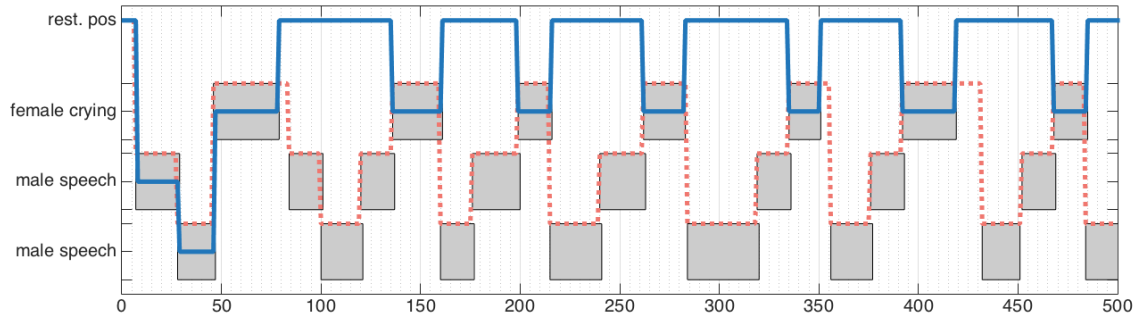
Une nouvelle fois, les durées de simulation T sont exprimées en pas de temps et correspondraient à un temps réel, sur le vrai robot, de $T \times 500$ms (une trame temporelle du *Blackboard* étant de 500 ms).

A chaque apparition d'un nouvel événement audiovisuel, le module DW va calculer sa Congruence et assigner un poids à l'objet correspondant. En fonction de ce poids, un mouvement de tête sera généré ou, au contraire, inhibé. La **Fig. 5.5** est une illustration des trois conditions de test générées : chaque rectangle gris représente le temps durant lequel une source audiovisuelle donnée émet un son, les traits bleus illustrent le comportement attentionnel du module DW et les traits rouges illustrent le comportement du robot naïf. Un rectangle traversé par un trait signifie donc que le système considéré a déclenché un ordre moteur vers l'objet concerné.

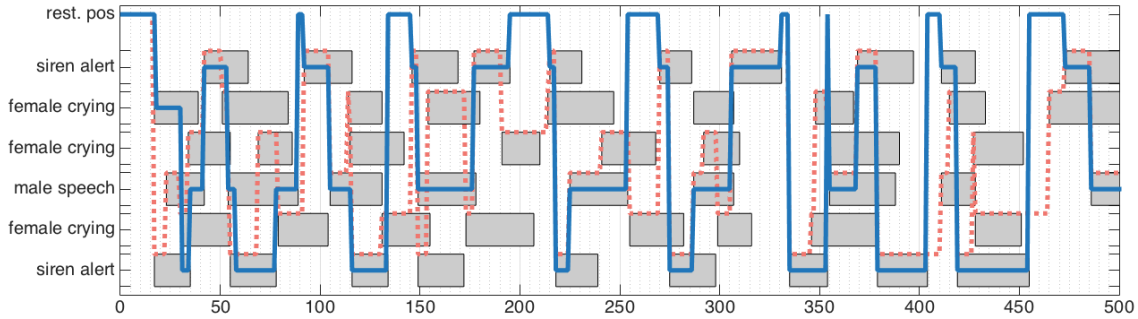
5.5.1.1 Objets focalisés

La figure **Fig. 5.5** nous permet tout d'abord d'observer la façon dont le module DW effectue son filtrage attentionnel sur la base de la Congruence des événements

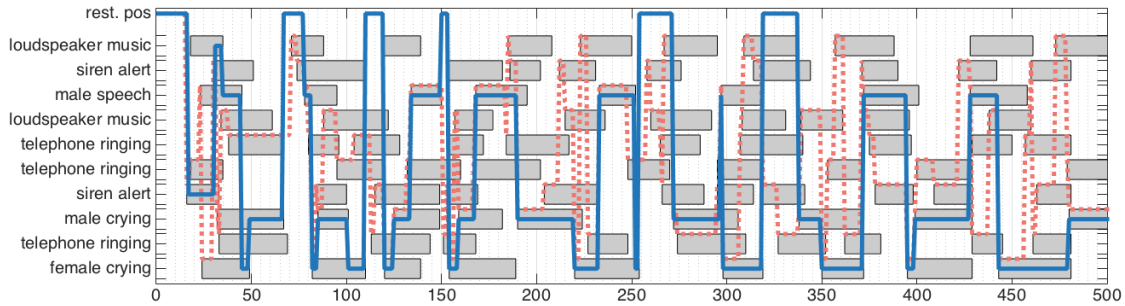
3. Se référer à la **Sec. 4.3** pour l'explication de ces notations.
4. Se référer au **Tab. 4.1** pour la liste des catégories audiovisuelles.



(a) Scénario n°1, 3 sources audiovisuelles



(b) Scénario n°2, 6 sources audiovisuelles



(c) Scénario n°3, 10 sources audiovisuelles

FIGURE 5.5 – OBJETS FOCALISÉS PAR LE DW — Sources audiovisuelles vers lesquelles un mouvement de tête est généré (*trait plein bleu*) par le module DW, (*trait pointillé rouge*) par le robot naïf. (*rectangles gris*) émission de son par la source audiovisuelle correspondante. Un objet est « focalisé » lorsque le trait traverse le rectangle.

audiovisuels apparaissant dans l'environnement. Dans le scénario n°1, cas unisource, les deux objets de catégorie *male speech* sont très rapidement ignorés au profit de $\mathcal{S}_1$ (*female crying*). Dans le scénario n°2, $\mathcal{S}_2$ (*female crying*) présente un intérêt pour le module DW en tout début d'exploration puis l'apparition de deux autres sources similaires ($\mathcal{S}_3$ et $\mathcal{S}_5$) rendent ce type d'événements de plus en plus congru, laissant ainsi la place pour d'autres types de sources. Le dernier scénario nous permet de mettre particulièrement en avant le mécanisme de sélection de la source audiovisuelle vers laquelle tourner sa tête. Dans ce scénario complexe à 10 sources audiovisuelles et jusqu'à 7 sources émettant simultanément, nous voyons que le module DW confère au robot une relative stabilité dans ses mouvements de tête, permettant notamment

d'ignorer complètement des sources, comme $\mathcal{S}_4$ (*loudspeaker music*) qui n'a jamais été la cible d'un mouvement de tête. Cette capacité à complètement ignorer une source audiovisuelle est un des intérêts majeurs du module DW. Nous pourrions considérer que les premières expériences du robot dans un nouvel environnement sont une forme de phase d'apprentissage. Bien que nous avons formalisé la notion de Congruence de telle façon qu'elle est appliquable même à partir de très peu de données, le comportement du module DW va connaître une forme de convergence lui permettant, une fois que l'environnement a été exploré (au sens des objets audiovisuels qui le composent), de pouvoir réagir instantanément à l'apparition de nouveaux événements.

L'environnement n°2 nous permet d'ailleurs d'observer un comportement particulier : le robot arrive à la conclusion que les sources d'intérêt sont pour lui, en autres, les sources $\mathcal{S}_1$ et $\mathcal{S}_6$ (*siren alert*). Or nous pourrions avancer que ce comportement n'est pas pertinent étant donné les sources présentes : trois *female crying* qui seraient sans doute plus prioritaires que la fixation des sirènes d'alarme. Cependant, cet exemple illustre deux points importants du module DW. Premièrement, la Congruence est une notion conférant à un objet audiovisuel une mesure de son éventuelle importance dans un environnement informationnel donné. Mais le comportement que nous, humains, sommes susceptibles d'appliquer dans des environnements tels le n°2 est également déclenché par de nombreux autres motivations qui se sont construites au fur et à mesure de notre expérience du monde. La Congruence est une première étape permettant d'effectuer un filtrage des objets audiovisuels perçus et ce, sans règles données *a priori*. Etant donné le scénario, le comportement du robot soumis au module DW ne sera donc dicté que par ce qu'il perçoit. De plus, lorsqu'il entre dans ce nouvel environnement, et particulièrement sans étape de transmission des connaissances, il est complètement naïf et n'a jamais été placé dans aucun environnement. Deuxièmement, le module DW (et nous reviendrons dessus plus tard) est une source de connaissance, similaire à celles décrites au **Sec. 3.2.3**, qui peut — et doit — être utilisée par d'autres sources du système TWO!EARS dans le but d'affiner la compréhension que le robot a de l'environnement. Cette connaissance pourrait aboutir au changement de la tâche à effectuer par le robot, étant donné l'apparition d'un événement d'importance requérant éventuellement une acquisition d'informations supplémentaires.

Enfin, en comparaison, le robot naïf (trait pointillé vert), n'étant motivé que par la Nouveauté d'une source sonore, tourne sa tête dès lors qu'une source émet un son. Ainsi, il ne réalise aucune analyse de contenu sémantique des événements apparaissant dans l'environnement : selon lui, ils sont tous égaux.

5.5.1.2 Mouvements de tête

La **Fig. 5.6** illustre le nombre de mouvements de tête générés par $\mathfrak{R}_n$ (flèches rouges) et par le module DW (flèches bleues) pour les trois conditions de test. Au centre de la figure se situe le robot. Les flèches pointent vers la position des sources sonores et leur longueur dénote le nombre de mouvements de tête qui ont été générés vers chacune des sources. Comme expliqué précédemment, nous avons placé toutes les sources sonores en face du robot afin qu'il ait toujours accès à la modalité visuelle.

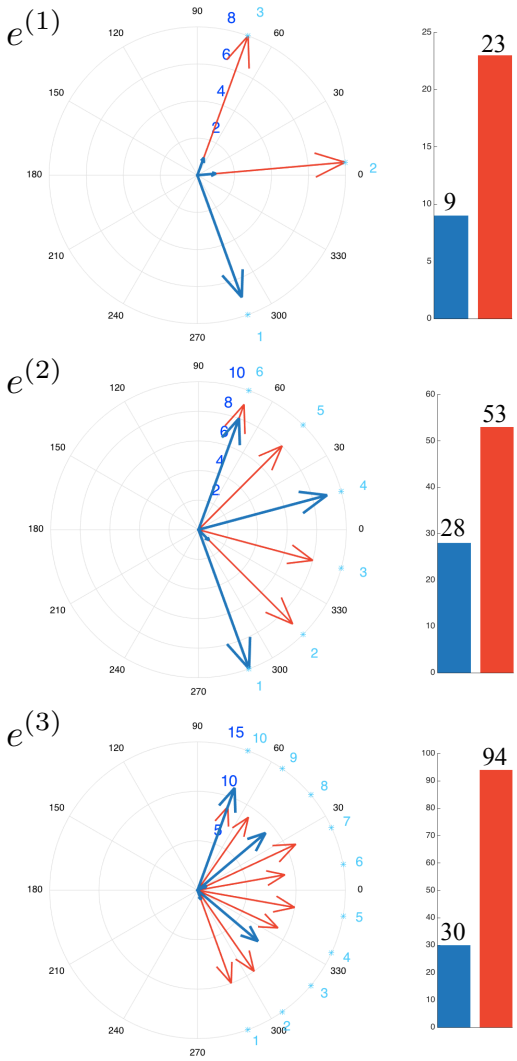
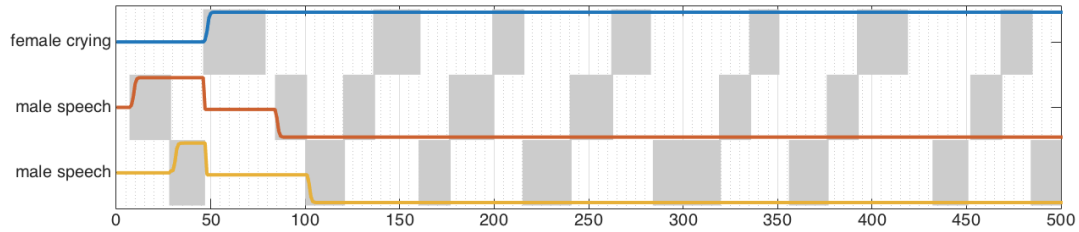


FIGURE 5.6 – MOUVEMENTS DE TÊTE EN CONDITIONS RÉELLES — llustration du nombre de mouvements de tête générés par (*bleu*) la HTMKS (*rouge*) le robot naïf virtuel. Chaque flèche pointe vers la position d'une source sonore et leur longueur dénote le nombre de mouvements vers la source pointée. (*histogrammes bleus*) nombre totaux de mouvements générés par (*violet foncé*) le module MFI, (*violet clair*) le module DW (les nombres en noir sont la somme des deux). (*histogrammes rouges*) nombre de mouvements générés par le robot naïf virtuel.

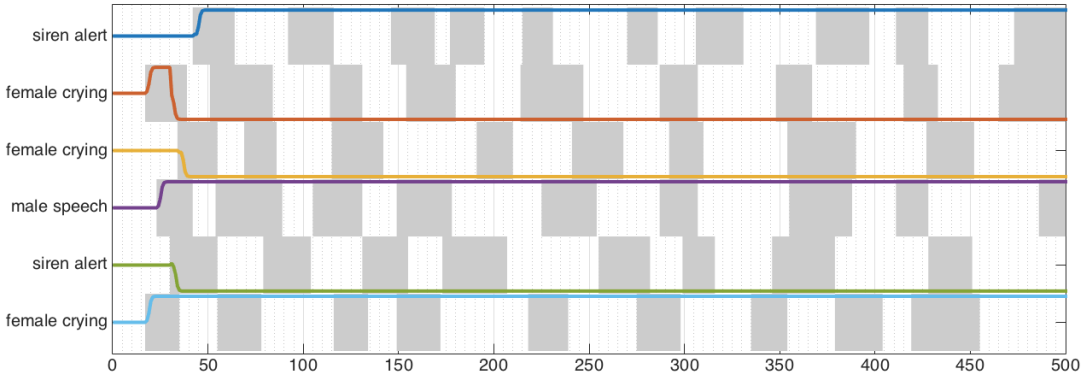
Deux observations peuvent être faites ici. Premièrement, le nombre de mouvements de tête est significativement inférieur dans le cas du module DW que dans celui du robot naïf. Nous réobservons le filtrage attentionnel effectué par le module DW sur la base du calcul de la Congruence de chaque objet. Deuxièmement, certaines sources ont été complètement ignorées : dans le cas du scénario n°2, aucun mouvement de tête n'a été généré vers $\mathcal{S}_3$ par exemple, source de catégorie *female crying*. En effet, étant donné l'état de l'exploration de l'environnement au moment où la $\mathcal{S}_3$ émet pour la première fois un son (à $t = 35$, cf. **Fig. 5.6**, (milieu)), le robot a déjà perçu deux autres sources de même catégories. Ainsi, le module DW juge cet événement comme *congru* et empêche ainsi, par diminution du poids de l'objet concerné, la génération d'un mouvement de tête vers cet objet.

5.5.1.3 Evolution des poids

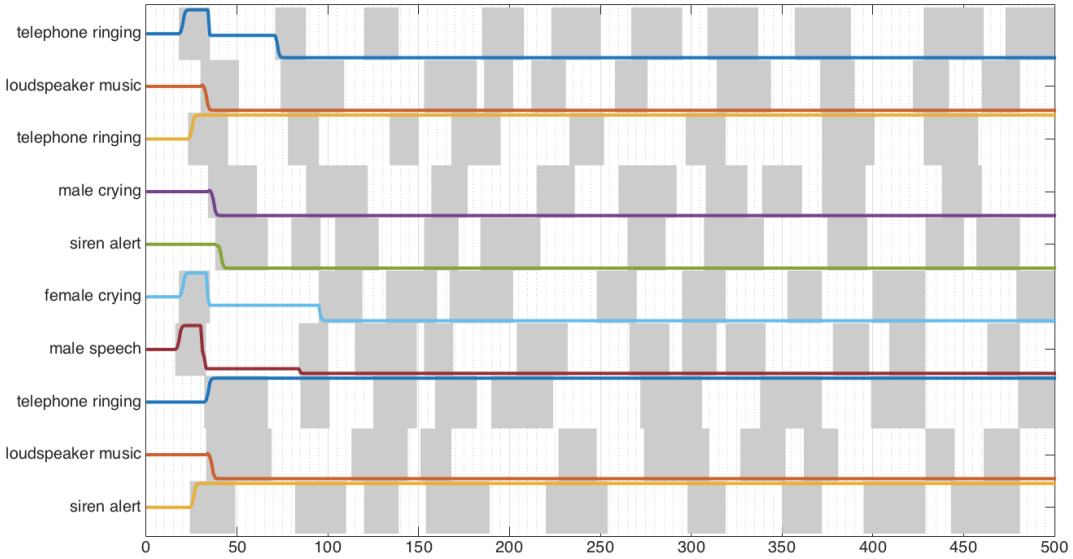
La **Fig. 5.7** illustre l'évolution des poids de chaque objet au cours de l'exploration de l'environnement. Dans le scénario n°1, l'apparition de $\mathcal{S}_3$ (*female crying*) affecte instantanément les poids des deux premières sources (*male speech*) qui étaient jusqu'à présent considérées comme incongrues (nous rappelons ici que les premières



(a) Scénario n°1, 3 sources audiovisuelles



(b) Scénario n°2, 6 sources audiovisuelles



(c) Scénario n°3, 10 sources audiovisuelles

FIGURE 5.7 – EVOLUTION DES POIDS, PAR SCÉNARIO — Chaque ligne de couleur représente le poids associé à un objet audiovisuel : (*position basse*) poids négatif et objet congru, (*position haute*) poids positif et objet incongru. Les mouvements de tête seront générés vers les objets ayant les poids les plus forts.

sources apparaissant dans l'environnement sont sujettes au cas limite imposé par l'inégalité non-stricte dans l'expression de $w(o_j)[t]$ de l'**Eq. 5.8** : elles seront définies comme incongrues tant qu'elles appartiendront à la même catégorie audiovisuelle). Le poids de $\mathcal{S}_1$ et $\mathcal{S}_2$ est réinitialisé et sera recalculé dès lors qu'elles émettront à

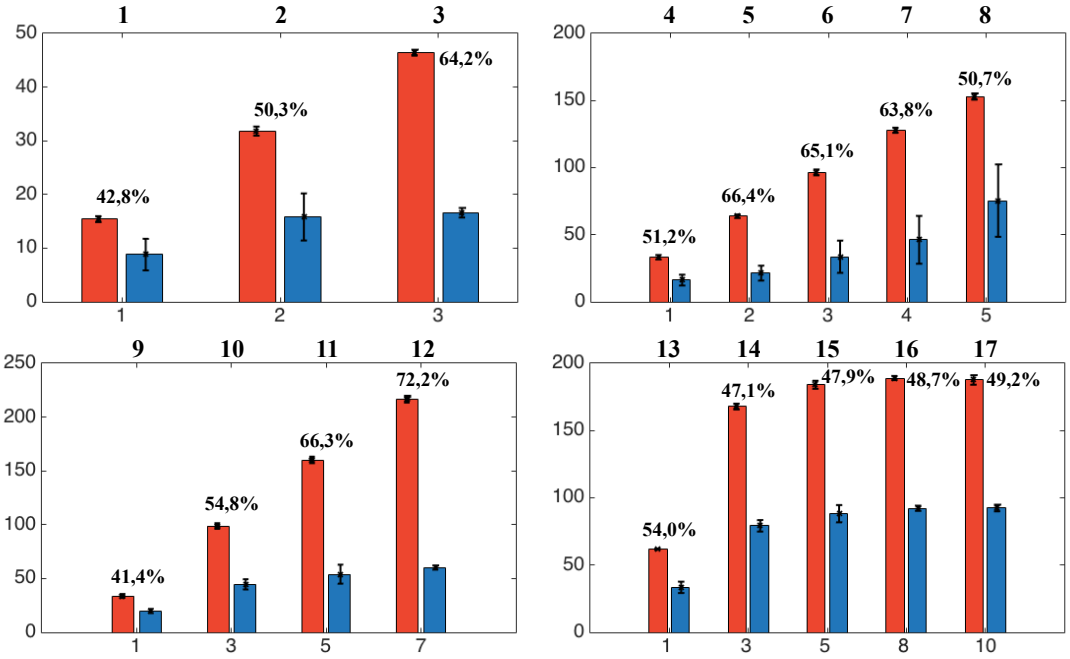


FIGURE 5.8 – NOMBRE DE MOUVEMENTS DE TÊTE GÉNÉRÉS — Histogramme des mouvements de tête générés par (*rouge*) le robot naïf et (*bleu*) le module DW. 17 conditions différentes ont été créées (se référer au texte pour leur description) pour cette évaluation, chaque condition ayant été testée cinq fois. Chaque histogramme représente la moyenne sur ces cinq tests, les barres d'erreurs représentant l'écart-type. Les pourcentages représentent le rapport entre les nombres de mouvements générés. Enfin, les numéros au-dessus des graphes correspondent au numéro de l'environnement simulé (cf. **Tab. 5.2**).

nouveau, à $t = 100$ et $t = 80$, respectivement. L'apparition de la $\mathcal{S}_3$ a eu pour conséquence de rendre $\mathcal{S}_1$ et $\mathcal{S}_2$ congrues à cet environnement. Jusqu'à la fin de cette exploration, puisqu'aucune nouvelle source n'apparaît, le module DW se focalise sur $\mathcal{S}_3$ exclusivement.

Nous observons le même type de comportement aux scénarios n°2 et n°3, cette fois-ci dans des environnements plus complexes. Notamment, nous observons au scénario n°3 que parmi les 10 sources audiovisuelles présentes, seules 3 sont détectées comme incongrues : $\mathcal{S}_3$, $\mathcal{S}_8$ et $\mathcal{S}_{10}$, dont deux appartiennent à la même catégorie. Cette réduction du nombre d'entités audiovisuelles susceptibles de présenter un intérêt pour le robot est de première importance car elle permet au robot de faciliter la compréhension d'environnements potentiellement complexes et de pouvoir interagir avec les entités présentes sur la base d'un calcul de Congruence.

5.5.2 Etude des mouvements de tête

Afin d'évaluer plus profondément l'impact du module DW sur les mouvements de tête, nous avons également effectué une série de tests en créant 17 environnements simulés différents en faisant varier (i) le nombre de sources et (ii) le nombre de sources

5. Se référer à la **Sec. 4.3** pour l'explication de ces notations.
6. Se référer au **Tab. 4.1** pour la liste des catégories audiovisuelles.

Conditions de test [5]					
$e^{(i)}$	n°	$n_{\mathcal{S}}$	n_{sim}^{max}	T	Catégories présentes [6]
$e^{(1)}$	1	3	1	500	1, 9
	2		2		
	3		3		
$e^{(2)}$	4	5	1	1000	1, 9, 18
	5		2		
	6		3		
	7		4		
	8		5		
$e^{(3)}$	9	7	1	1000	1, 9, 18, 21, 28
	10		3		
	11		5		
	12		7		
$e^{(4)}$	13	10	1	1000	1, 3, 9, 18, 21, 28, 36
	14		3		
	15		5		
	16		8		
	17		10		

TABLE 5.2 – Caractéristiques des 4 environnements pour lesquels 17 scénarios ont été générés pour étudier le nombre de mouvements de tête générés par le module DW. Ces scénarios sont de complexité croissante, tant au niveau du nombre de sources $n_{\mathcal{S}}$ que du nombre maximum de sources émettant un son simultanément n_{sim}^{max}. Comme précédemment, ces résultats seront comparés aux performances du robot naïf.

simultanées. Ces 17 conditions sont présentées au **Tab. 5.2**. Toutes les conditions ont été testées cinq fois, chaque fois avec un décours temporel différent et aléatoire, du point de vue de l'assignation des catégories audiovisuelles aux sources données.

La **Fig. 5.8** résume l'ensemble des résultats obtenus sur ces 17 conditions et 85 simulations. Chaque histogramme est la moyenne pour chaque condition sur les cinq tests effectués pour chacune d'entre elles. Les histogrammes rouges représentent le nombre d'ordres moteurs générés par le robot naïf tandis que les histogrammes bleu foncé représentent ceux générés par le module DW. Enfin, les barres d'erreur représentent les écarts-types et les pourcentages sont le rapport entre les nombres de mouvements générés par le robot naïf et celui soumis au DW.

Ces résultats montrent une diminution systématique et conséquente des mouvements de tête générés par le module DW. Dans ces quatre scénarios allant du plus simple au plus complexe, ces mouvements ont été diminués de $52,5\%$, $59,5\%$, $58,7\%$ et $49,4\%$ respectivement, pour une moyenne totale de $55,0\%$. Ainsi, en moyenne, la moitié des mouvements de tête sont filtrés par le module DW en comparaison du robot naïf. Nous verrons plus tard que, bien que cette diminution soit déjà conséquente, le module DW pourrait être plus performant dans son inhibition des mouvements de tête, c'est-à-dire dans le choix des sources audiovisuelles nécessitant l'attention du robot.

5.5.3 Différents environnements

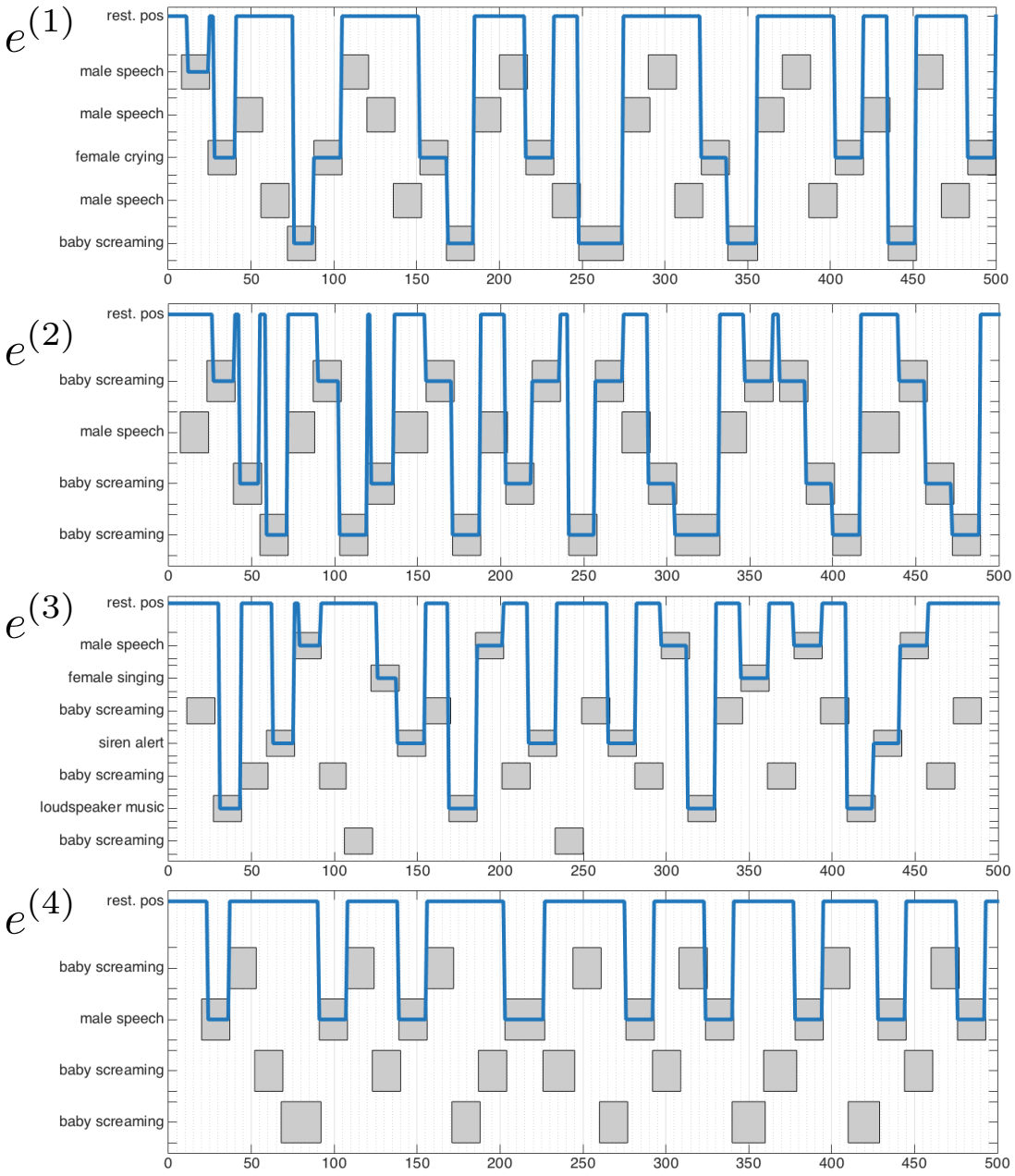


FIGURE 5.9 – OBJETS FOCALISÉS PAR LE DW & TRANSMISSION DES CONNAISSANCES — Sources audiovisuelles vers lesquelles un mouvement de tête est généré par le module DW (pour des soucis de clarté, le robot naïf tournant sa tête chaque fois qu'une source émet n'a pas été représenté). (*rectangles gris*) émission de son par la source audiovisuelle correspondante. Un objet est « focalisé » lorsque le trait traverse le rectangle. Chaque environnement a été exploré successivement permettant d'observer la transmission des connaissances d'un environnement à l'autre.

Un des intérêts majeurs d'« apprendre » est de pouvoir réutiliser les connaissances aquises dans de nouvelles situations. Dans le cadre du module DW, la transmission des connaissances se fait lorsque le système considère qu'un nouvel environnement inconnu en cours d'exploration est le même qu'un environnement déjà

exploré précédemment. Lors de l'exploration d'un nouvel environnement, et s'appuyant sur notre définition d'un ENVIRONNEMENT (cf. **Déf. 5**), le module DW va déterminer, à chaque apparition d'un nouvel objet, s'il peut appliquer les règles de Congruence apprises dans un des environnements précédemment explorés. Cette méthode permettrait ainsi d'accélérer le processus de réaction attentionnelle en appliquant dès les premières occurrences d'événements audiovisuels dans un environnement inconnu un ensemble de règles comportementales, formalisées par la notion de Congruence.

Pour tester cette transmission de connaissances, nous avons défini un scénario évolutif permettant d'observer l'impact de l'exploration de plusieurs environnements sur le comportement du module DW. Le **Tab. 5.3** détaille les caractéristiques des quatre environnements simulés.

Conditions de test 4.6 [7]				
$e^{(i)}$	$n_{\mathcal{S}}$	n_{sim}^{max}	T	Catégories présentes [8]
1	5	3	300	1, 9, 18
2	3	1	300	1, 18
3	7	4	300	1, 11, 18, 41
4	3	1	300	1, 18

TABLE 5.3 – Caractéristiques des 4 environnements générés pour étudier la capacité du module DW à utiliser les connaissances apprises pour l'exploration de nouveaux environnements inconnus.

Ce scénario va se dérouler de la façon suivante :

1. le robot explore un premier environnement $e^{(1)}$, il est donc complètement ignorant et établit ses premières règles de Congruence ;
2. le robot explore un deuxième environnement $e^{(2)}$ constitué d'objets audiovisuels déjà rencontrés dans le premier environnement. Ainsi, il va pouvoir appliquer les règles de Congruence apprises de $e^{(1)}$;
3. le robot explore maintenant un troisième environnement $e^{(3)}$ contenant un objet appartenant à une nouvelle catégorie audiovisuelle. Le module DW va ainsi créer un nouvel environnement et apprendre de nouvelles règles de Congruence ;
4. le robot explore un dernier environnement $e^{(4)}$, identique à $e^{(2)}$. L'exploration de $e^{(3)}$ va avoir un impact sur le comportement du robot dans ce dernier environnement.

La **Fig. 5.9** illustre les objets focalisés dans chacun des environnements successivement explorés par le robot soumis seulement au module DW. Le premier environnement est exploré de façon classique par le module DW : les sources $\mathcal{S}_3$ (*female crying*) et $\mathcal{S}_5$ (*baby screaming*) sont considérés, en fin de simulation, comme des événements incongrus tandis que les autres sources sont, elles, ignorées car de catégories considérées comme congrues à l'environnement $e^{(1)}$. L'impact sur l'exploration de l'environnement $e^{(2)}$ est évident : malgré le rapport entre le nombre d'objets de catégories *baby screaming* et *male speech* (trois contre un), le module DW considère

7. Se référer à la **Sec. 4.3** pour l'explication de ces notations.
8. Se référer au **Tab. 4.1** pour la liste des catégories audiovisuelles.

qu'étant donné son expérience passée, cet environnement est susceptible de ressembler au tout premier environnement exploré et génère ainsi des mouvements de tête vers les sources $\mathcal{S}_1$, $\mathcal{S}_3$ et $\mathcal{S}_4$. Si cet environnement avait été exploré en premier, les mouvements de tête auraient été exactement inverses : ces trois sources se seraient vues ignorées très rapidement afin de ne se focaliser que sur la source $\mathcal{S}_2$ (*male speech*).

Dans l'environnement $e^{(3)}$, la première source émettant étant de catégorie *baby screaming*, le module DW juge que cet environnement est susceptible d'être le même que $e^{(2)}$ et donc que $e^{(1)}$. Les règles de Congruence $\mathbf{W}$ appliquées sont donc $\mathbf{W}^1$, de l'environnement $e^{(1)}$. Cependant, l'apparition de la source $\mathcal{S}_6$ (*loudspeaker music*) appartenant à une toute nouvelle catégorie audiovisuelle entraîne la création d'un nouvel environnement avec ses propres règles de Congruence. L'inhibition consécutive d'un mouvement de tête vers $\mathcal{S}_5$ (*baby screaming*) est ainsi due à la Congruence de cet événement étant donné l'environnement $e^{(3)}$. A $t = 75$, la catégorie de la source $\mathcal{S}_1$ (*male speech*) aurait été considérée comme congrue si les règles de Congruence de $e^{(1)}$ avaient été appliquées, comme au début de l'exploration de l'environnement actuel. En revanche, nous voyons qu'un mouvement de tête est déclenché en raison de l'incongruence de cet événement étant donné l'environnement $e^{(3)}$. Le module DW est donc bien en train d'apprendre un tout nouvel environnement possédant ses propres règles de Congruence.

Le dernier environnement $e^{(4)}$ permet d'observer l'impact de l'exploration de $e^{(3)}$ sur l'exploration d'un environnement nouveau mais identique à $e^{(2)}$. Pour rappel, dans $e^{(2)}$, les événements appartenant aux catégories *baby screaming* étaient considérés comme incongrus par application des connaissances issues de $e^{(1)}$. Or nous voyons dans $e^{(4)}$ que l'inverse se produit : la source $\mathcal{S}_2$, de catégorie *male speech*, est la seule à être focalisée par le module DW, toutes les autres sources étant tout simplement ignorées. De plus, nous observons l'inhibition complète et dès le début de l'exploration des mouvements de tête vers les sources $\mathcal{S}_1$, $\mathcal{S}_3$ et $\mathcal{S}_4$: cette transmission des connaissances permet ainsi d'accélérer l'analyse de l'environnement puisque le robot adopte déjà un comportement issu de son expérience passée dans un environnement inconnu.

5.5.4 Discussion

Rappelant les très nombreux travaux de recherche menés sur la saillance, dont certains ont été exposés à la **Sec. 2.3.1.2**, nous pourrions dire que chaque nouvelle émission d'un son par une source audiovisuelle constitue un événement saillant. Les sources déjà présentes devenant, progressivement, un bruit de fond, lorsqu'une source se met à produire un son, cette nouveauté représente une saillance dans l'environnement audiovisuel actuel. Basé sur cette considération, nous avons défini un robot naïf tournant sa tête à chaque fois qu'un nouvel événement survient dans l'environnement. Ce robot, dont nous pouvons dire qu'il est motivé par la *Nouveauté* ou la *Curiosité*, a été notre référence en tant que système attentionnel ne possédant pas non plus de règles de comportement données *a priori* : il s'agit d'un robot purement réactif privé de capacités d'analyses avancées des événements audiovisuels apparaissant au cours de l'exploration.

De l'autre côté, nous avons testé un robot soumis au *Dynamic Weighting* tournant sa tête à chaque fois qu'un nouvel événement survenant est incongru à l'environnement en cours d'exploration. L'analyse de la Congruence d'un objet audiovisuel permet de ne pas avoir à donner de règles *a priori* sur quel contenu sémantique doit attirer l'attention du robot.

Les premiers résultats obtenus ici nous ont permis d'observer la façon dont le module DW analyse la scène audiovisuelle et le filtrage attentionnel dont il est capable, sur la base du calcul de la Congruence. A chaque fois, le nombre de mouvements de tête générés a été significativement inférieur au nombre de mouvements générés par le robot naïf. De plus, la capacité du robot à complètement ignorer certaines sources émettant pour la première fois présente un avantage considérable en cela qu'il permet de donner la possibilité au robot de se concentrer sur d'autres sources sonores éventuellement d'intérêt. Ainsi, un comportement exploratoire, par l'entremise de rotations de la tête, dirigé par le module *Dynamic Weighting* permet d'ajouter un niveau supplémentaire dans la compréhension de la scène audiovisuelle. Ce niveau de compréhension ne prend en compte que les caractéristiques sémantiques des données audiovisuelles et constitue ainsi une nouvelle source de connaissance utilisable par d'autres KS *via* le Blackboard.

5.6 Conclusion du Chapitre

Le module DW a été le premier composant du modèle HTM. Le problème qu'il a tenté de résoudre, avec les contraintes fortes posées par Two!Ears, était de déterminer quel événement audiovisuel nécessite l'attention du robot, dans un contexte d'exploration d'environnements inconnus. Le module DW constitue une source de connaissance supplémentaire sur les objets audiovisuels qui composent un environnement. La définition de la notion de *Congruence* a permis l'élaboration du module DW. Celle-ci peut être vue comme une mesure de l'importance relative à moyen terme d'un événement audiovisuel perçu au sein d'un environnement en cours d'exploration. Le parti pris a donc été de ne pas rechercher ce qui définit la Congruence dans les caractéristiques bas niveau des signaux perçus mais au contraire se placer plutôt au niveau sémantique, c'est-à-dire après l'analyse des différents experts d'identification audiovisuels. Ce choix de se situer loin du signal brut a aussi été motivé par le fait que la vision et l'audition sont deux sens très différents mais fonctionnant de façon très complémentaire. Le système auditif humain, bien qu'extrêmement puissant, bénéficie d'une collaboration avec le système visuel en cela qu'il permet le déclenchement de mouvements de tête aboutissant à une représentation multimodale de l'objet d'intérêt. La vision, étant, par nature, beaucoup plus précise que l'audition (ne serait-ce que parce que dans une image les flux perceptifs sont généralement très distinguables, contrairement à l'audition), l'apport des informations visuelles permet une analyse plus fine et précise de l'objet, que ce soit son identité ou sa localisation. L'audition est donc ici utilisée comme un « signal d'alarme » envoyant une requête de mouvement de tête.

Mais les résultats de l'analyse du module DW ne sont pas une fin en soi. En effet, réduire un comportement attentionnel robotique au calcul de la Congruence des événements audiovisuels n'est pas suffisant pour doter un robot exploratoire

d'une capacité de compréhension globale de l'environnement pertinente. Nous avons considéré le module DW comme un système permettant de réduire la complexité des données en entrée en cela qu'il offre une analyse sémantique de l'environnement et fournit au Blackboard une couche supplémentaire d'interprétation qui pourra être utilisée par d'autres KS. La Congruence d'un objet est donc une indication sur quels objets *pourraient* nécessiter l'attention du robot. Cette indication devrait être incluse dans un ensemble d'autres analyses, plus ou moins haut niveau, permettant, *in fine*, de doter le robot d'une véritable capacité à analyser intelligemment son environnement.

Notamment, dans le scénario le plus complexe de Two!Ears de S&R, les indications fournies par le module DW pourraient être incluses par une KS intégrant le résultat des analyses de toutes les KS dédiées à l'analyse sémantique de la scène auditive, permettant ainsi de décider si la source audiovisuelle jugée comme d'intérêt par le module DW l'est suffisamment en regard de la tâche de sauvetage à accomplir. Par exemple, les objets audiovisuels de type *siren alert* pourraient être considérées comme non pertinentes dans une tâche de sauvetage, tandis que les sources de type *baby screaming* pourraient nécessiter un mouvement de tête. Ce mouvement de tête sera alors utilisé pour avoir plus d'informations sur ce nouvel objet audiovisuel aboutissant éventuellement en une redéfinition de la tâche de sauvetage : sauver un bébé pourrait être défini comme une tâche plus prioritaire que celle de sauver un adulte (le conditionnel est ici important).

Le module DW a été motivé par de nombreuses considérations biologiques comme le fonctionnement des aires perceptives face à des événements imprédictibles (*Mismatch Negativity* ou réseaux neuronaux impliqués dans l'attention et dans sa réorientation), la boucle ganglions de la base — thalamus — cortex et modèle GPR (sélection des actions motrices), le calcul de probabilité *a posteriori* et l'inclusion d'une forme de probabilité conditionnelle entre les différentes catégories audiovisuelles détectées par le robot (considérations se rapprochant du paradigme bayésien largement observé dans les processus cérébraux et perceptifs en particulier) ou encore le fonctionnement du colliculus supérieur (intégration multimodale et génération consécutive de mouvements de tête, entre autres). De ces inspirations de l'étude des mécanismes cérébraux — et seulement une inspiration, non une tentative de mimétisme — a émergée la tentative de réaliser un algorithme d'analyse d'une scène audiovisuelle simple mais laissant le plus possible au robot la faculté d'apprendre sans données *a priori*. La seule connaissance dont ce module a besoin est la classification audiovisuelle des sources présentes, ainsi que leur localisation, analyses effectuées ici par des KS dédiées.

Cependant, un problème majeur apparaît ici, lié aux données auxquelles le robot a accès.

Dans le cas où un objet est situé *derrière* le robot, le module DW, tel qu'il est implémenté, aura préalablement besoin de tourner la tête vers cet objet afin d'obtenir le label visuel manquant et de pouvoir calculer sa Congruence... pour éventuellement requérir la génération d'un mouvement de tête vers cet objet... Tout l'intérêt de la modulation des mouvements de tête (génération ou inhibition) est mis en cause. C'est la raison pour laquelle nous avons choisi, dans la partie résultats, de simuler un environnement dans lequel toutes les sources audiovisuelles sont situées dans le

champ de vision (artificiellement large) du robot. Ainsi, nous avons permis au robot d'avoir toutes les données disponibles afin qu'il puisse calculer la Congruence des événements audiovisuels apparaissant dans cet environnement.

Afin de résoudre ce problème majeur (dans un environnement réaliste, le champ de vision du robot est beaucoup plus petit et les sources sonores sont possiblement réparties partout autour de lui), il a été nécessaire de coupler au module DW un deuxième module : le module *Multimodal Fusion & Inference* (module MFI). Ce module a pour but d'apprendre à inférer d'éventuelles données manquantes et ce, en respectant les mêmes contraintes que le module DW, à savoir un apprentissage en ligne et en temps réel et sans règles données *a priori*. Ce deuxième module se situe juste avant le module DW lui permettant ainsi de toujours avoir accès à une représentation multimodale des événements qu'il a à traiter.

Le chapitre suivant est dédié à la description de ce module.

Chapitre 6

Module d'Inférence et de Fusion Multimodale

Q. Comment apprendre à compléter une information éventuellement manquante sur un objet audiovisuel sur la seule base des données perçues ?

Le module d'Inférence et de Fusion Multimodale (*Multimodal Fusion & Inference*, MFI) bien qu'implémenté après le module DW est situé, dans la chaîne de traitement des données perçues, juste avant lui. En effet, nous avons vu précédemment que le module DW possède un défaut majeur : il fonctionne sur la base d'objets *audiovisuels*, impliquant donc la connaissance des labels audio et visuels. Mais lorsque l'objet est situé derrière le robot, il est impossible au module DW d'avoir accès à l'information visuelle et ainsi de calculer la *Congruence*. De plus, les experts d'identification sont sujets à erreurs quant à la classe audio ou visuelle à laquelle un événement appartient. Ainsi, la décision prise par le module DW sera aussi potentiellement erronée. Ces deux sources d'erreurs possibles ont en commun le fait qu'elles vont avoir un impact sur la fusion des labels audio et visuels et ainsi perturber le fonctionnement du module DW.

Le module MFI va ainsi tenter de résoudre ce problème en proposant un algorithme d'apprentissage basé sur la capacité du robot à tourner sa tête, mouvement qui sera utilisé pour apprendre le lien entre la modalité audio et la modalité visuelle. Le module MFI a donc deux objectifs principaux :

1. apprendre le lien entre la modalité audio et la modalité visuelle ;
2. corriger les erreurs éventuellement présentes dans les sorties des Knowledge Sources sur lesquelles il se base.

La **Fig. 6.1** montre la structure interne du module MFI et particulièrement les deux parties qui le composent :

- le réseau de neurones artificiel responsable de la catégorisation correcte de la énième trame audiovisuelle (partie de gauche) ;
- l'ordre moteur généré pour confirmer ou infirmer l'inférence effectuée par le module MFI lorsqu'une modalité est manquante (partie de droite).

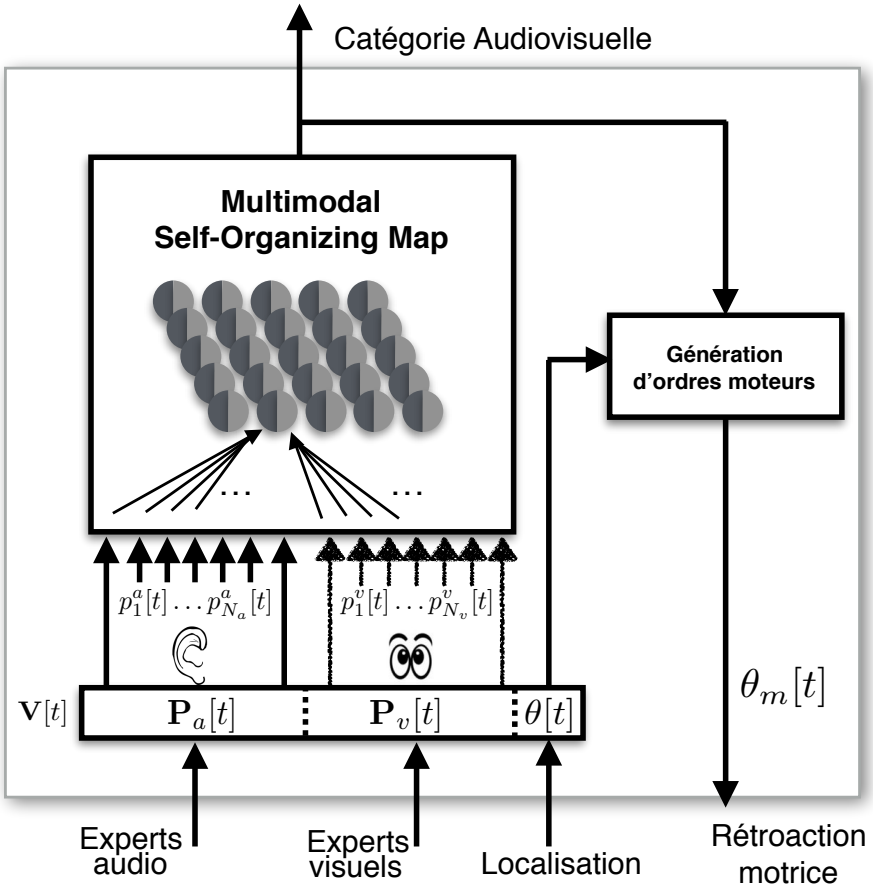


FIGURE 6.1 – ARCHITECTURE GLOBALE DU MODULE MFI — Les données issues des experts d'identification sont envoyées au *Multimodal Self-Organizing Map* (**Sec. 6.2**) pour l'apprentissage du lien entre la modalité audio et la modalité visuelle. Le module MFI est également responsable de la génération de mouvements de tête, sur la base de la sortie de l'expert de localisation audio. Le module MFI a pour but d'estimer la catégorie audiovisuelle de l'objet perçu au temps t.

Ainsi, le module MFI peut être considéré comme un système de fusion de classifieurs *actif* qui estime la catégorie audiovisuelle $\widehat{\mathcal{C}}[n]$ d'un objet perçu. La deuxième partie est en charge de la prise de décision sur les mouvements de tête : si une modalité est manquante ou si la confiance en l'estimation de la catégorie $\widehat{\mathcal{C}}[n]$ est trop faible, un mouvement sera requis dans le but de confirmer — et donc renforcer — ou infirmer l'inférence.

Ce chapitre est organisé comme suit :

La section **Sec. 6.1** introduit et décrit le fonctionnement des cartes auto-organisatrices, ou *Self-Organizing Map* (SOM). C'est à partir de cet algorithme d'apprentissage que nous avons développé le *Multimodal-Self Organizing Map*.

La section **Sec. 6.2** décrira le fonctionnement du *Multimodal Self-Organizing Map* (M-SOM), adaptation d'une carte auto-adaptative. Le M-SOM est le centre du module MFI : c'est cette structure qui est en charge de l'apprentissage des modalités audio et visuelle.

La section **Sec. 6.3** décrit la façon dont le module MFI va réaliser la double fusion

des données issues des classifieurs : au sein de chaque modalité puis entre les modalités.

La **Sec. 6.4** enfin, décrit la façon dont les ordres moteurs peuvent être générés par le module MFI.

6.1 Carte auto-adaptative classique

Nous avions, à la **Sec. 2.4.1.1**, justifié notre choix d'une carte auto-organisatrice (ou SOM) pour effectuer la partie d'apprentissage nécessaire au modèle HTM. Nous avions également mis en avant le fait que cet algorithme nécessiterait d'être modifié pour devenir utilisable dans le cadre du problème au sein duquel nous nous situons. Cette section consiste donc en la description préalable de l'algorithme SOM classique, base sur laquelle nous avons développé le *Multimodal Self-Organizing Map*, décrit à la suite de la section suivante.

6.1.1 Formalisation

Un SOM est une carte en deux dimensions composée de $I \times J$ nœuds [1] interconnectés, que nous notons $r_{i,j}$. A chaque nœud est associé :

1. un vecteur de poids $\mathbf{w}_{i,j}$ de la même dimension que les données d'entrée ;
2. une position (i, j) dans l'espace de la carte ;
3. des connexions $\chi_{(ij)\to(kl)}$, où $[i, k] \in [1, I]$ et $[j, l] \in [1, J]$, entre le neurone $r_{i,j}$ et ses neurones voisins, avec une exception pour les neurones présents aux bords qui possèdent moins de voisins que les autres. A noter que cette connectivité peut être modifiée en fonction de l'application du SOM.

Ce sont les vecteurs de poids $\mathbf{w}_{i,j}$ associés aux neurones $r_{i,j}$ qui vont coder l'espace des données en entrée.

Dans tout ce qui suit, nous considérerons, comme exemple, une matrice de données à apprendre notée $\mathbf{P}$, de taille $(M \times N)$, avec M étant le nombre d'observations des N caractéristiques des données d'entrée. Le SOM va donc apprendre à représenter les données multidimensionnelles d'entrée de la matrice $\mathbf{P}$ dans un espace à deux dimensions : à chaque catégorie de données détectée par le SOM va correspondre un ou plusieurs nœuds de la carte.

L'algorithme SOM, comme de la plupart des algorithmes d'apprentissage, est itératif. Une itération d'apprentissage va consister en trois étapes : (i) la sélection aléatoire d'un vecteur $\mathbf{p} \in \mathbf{P}$, de taille N, (ii) la recherche du nœud $r_{i,j}$ dont le vecteur de poids associé $w_{i,j}$ est le plus *ressemblant* au vecteur d'entrée $\mathbf{p}$, (iii) le renforcement de cette ressemblance pour le nœud concerné puis la propagation de cette ressemblance aux voisins de ce nœud. Ce processus est répété jusqu'à ce que les M vecteurs de la matrice $\mathbf{P}$ aient été utilisés. Ensuite, une nouvelle itération démarre. Les itérations d'apprentissage ne correspondent pas simplement en la répétition des trois étapes mentionnées ci-dessus. A chaque itération, la façon dont la

1. Aussi appelés neurones, toujours en référence à l'inspiration biologique.

ressemblance est renforcée et propagée est modifiée (détails ci-après). Cela permet, entre autres, de faire converger l'apprentissage.

La notion de ressemblance est formalisée par la *distance euclidienne* entre le vecteur d'entrée et chacun des vecteur de poids : le nœud le plus ressemblant sera donc le nœud possédant un vecteur de poids ayant la *distance euclidienne la plus petite* entre lui et le vecteur d'entrée. Ce nœud « gagnant » est appelé la *Best Matching Unit*[2] (BMU) du vecteur d'entrée. C'est ce nœud, étant le meilleur représentant du vecteur d'entrée, qui va servir de point de départ, du point de vue spatial, de l'apprentissage à proprement parler du réseau. A la fin de l'apprentissage, chaque nœud $r_{i,j}$ représentera un point particulier de l'espace d'entrée. En reprenant l'exemple de la cochlée, la carte de Kohonen résultant de l'apprentissage d'une suite de stimuli sonores à des sinus purs contiendrait des zones bien déterminées codant chaque fréquence fondamentale observée.

Dans ce qui suit, une itération d'apprentissage sera notée $t \in [1, \ldots, N_{it}]$, où N_{it} est le nombre d'itérations total.

Initialisation du réseau. La première étape de l'utilisation d'un SOM consiste en l'initialisation des vecteurs de poids de chaque neurone. Chaque composante $\mathbf{w}_{i,j}^n$ des vecteurs de poids sera initialisée selon $\mathbf{w}_{i,j} \in \mathbb{R}^+ \rightarrow [0, 1]$. Cette initialisation peut être aléatoire ou linéaire [265]. La méthode aléatoire consiste en l'assignation de valeurs choisies au hasard, limitant l'introduction d'un quelconque biais dans l'organisation du réseau avant l'apprentissage. La méthode linéaire quant à elle, permet d'accélérer la convergence de l'apprentissage en initialisant les vecteurs de poids après une étape d'Analyse en Composantes Principales (ACP[3]) de la matrice d'entrée à apprendre. Nous avons choisi la méthode alétoire puisque nous ne donnons pas au système de connaissance sur la distribution ou l'organisation des données, préalablement à l'exploration.

Recherche de la BMU. A l'itération t, il s'agit de trouver le nœud r_{BMU} dont le vecteur de poids associé $\mathbf{w}_{i,j}$ est le plus proche du vecteur d'entrée $\mathbf{p}$, selon :

$$r_{\text{BMU}} = r_{i,j}[t], \text{ avec } (i,j) = \arg\min_{(i,j)}\{\|\mathbf{p} - \mathbf{w}_{ij}\|\} \tag{6.1}$$

où $\|.\|$ représente la distance euclidienne.

Mise à jour. Une fois la BMU trouvée, les vecteurs de poids de chaque nœud sont mis à jour à l'aide d'une fonction de voisinage $h_{i,j}$ permettant la propagation de l'apprentissage et la création de la topologie autour de cette BMU :

2. Unité de plus forte ressemblance.

3. L'Analyse en Composantes Principales est une méthode de réduction de l'espace d'entrée par détermination des variables d'observation participant le plus à la variabilité des données. Les variables les moins explicatives pourront ainsi être retirées sans altérer le comportement global des données.

$$\mathbf{w}_{ij}[t+1] = \mathbf{w}_{ij}[t] + \alpha[t]\ h_{ij}[t]\ \|\mathbf{p} - \mathbf{w}_{ij}[t]\| \tag{6.2}$$

où $h \to \mathbb{R}$ étant la fonction de voisinage gaussienne avec :

$$h_{ij}[t] = \exp\left(-\frac{\|r_{\text{BMU}}[t] - r_{ij}\|^2}{2\sigma[t]^2}\right) \tag{6.3}$$

où σ est la variance de la fonction gaussienne, paramètre qui aura un impact sur l'amplitude de la propagation de la ressemblance entre le vecteur d'entrée et la BMU.

L'**Eq. 6.3** montre que la fonction de voisinage est dépendante de l'itération t : plus l'itération est élevée, i.e. plus l'apprentissage est avancé, moins la fonction de voisinage est ample, impliquant donc une réduction de la taille de la zone touchée par la propagation de la ressemblance entre la BMU et le vecteur $\mathbf{p}$. Une fois les étapes de recherche de BMU et de mise à jour des poids effectuées, un autre vecteur $\mathbf{p}$ de la matrice $\mathbf{P}$ est choisi et ces mêmes étapes sont reconduites jusqu'à ce que toute la matrice ait été explorée. Et une fois toute la matrice explorée, l'ensemble du processus (à l'exception de l'initialisation du réseau) sont répétées T fois. Une fois la phase d'apprentissage terminée, le SOM peut être utilisé pour effectuer une catégorisation des données apprises ou de données nouvelles.

Utilisation du SOM. Une fois que le réseau a appris les données qui lui ont été fournies, le réseau peut être utilisé à deux fins. La première est de récupérer les catégories créées par le SOM afin d'étudier la façon dont les données sont liées et corrélées entre elles. Cette étape de catégorisation est essentielle : c'est le but du SOM de pouvoir réduire des données de haute dimension (nombre M de composantes des vecteurs d'entrée) en une représentation de dimension beaucoup plus faible, deux dans notre cas (position dans la carte). De plus, les ensembles de nœuds de la carte codant une information similaire seront vus comme des catégories ce qui permet de considérer également une réduction de la dimension de la matrice d'entrée $\mathbf{P}$ entière : d'un nombre d'exemples N, le SOM les réduit à leur appartenance à N_c groupes, ou catégories.

Ainsi, le SOM peut être utilisé (i) afin de diviser un ensemble de données en catégories et (ii) dans le but de déterminer à quelle catégorie appartient un vecteur d'entrée unique, appris ou nouveau, en analysant à quelle catégorie de neurones la BMU de ce vecteur appartient.

Preuve de convergence. Comme pour tout algorithme de *machine learning*, la convergence de l'apprentissage doit pouvoir être prouvée (même si dans certains cas, comme celui des réseaux de neurones profonds, cette convergence est difficilement calculable). Dans le cas de l'algorithme SOM, plusieurs méthodes permettent de juger si la carte est stable, robuste et pertinente. Nous citerons notamment le calcul de la U-Matrix [266, 267] : chaque nœud est caractérisé par sa distance aux neurones voisins (voir paragraphe suivant pour quelques exemples de voisinage) et représentée sous forme de nuances de gris (cf. **Fig. 6.2**). La distance utilisée doit être la même

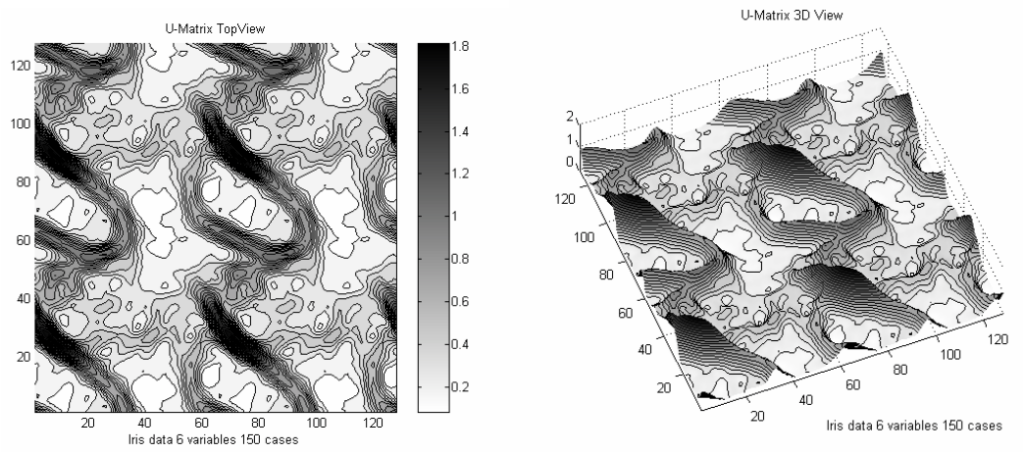


FIGURE 6.2 – U-MATRIX — Représentation en nuances de gris de la distance entre chaque neurone et ses neurones voisins selon la méthode de « U-Matrix » développée par ALFRED ULTSCH [266]. Plus la nuance de gris est claire, plus la distance est faible. Les données employées ici sont issues des données Iris de Fisher [268] (figure d'après [267]).

que celle employée lors de l'apprentissage, euclidienne dans notre cas. Cette représentation permet de visualiser l'organisation du réseau et de voir ainsi les différentes catégories que le SOM a détectées au sein de la matrice de données d'entrée. Un SOM correctement paramétré et ayant convergé aboutira à une carte aux zones bien différenciées et dont les frontières ne seront pas changées par l'apprentissage d'un nouvel exemple de la matrice d'apprentissage.

Il est intéressant de noter la publication toute récente d'une nouvelle mesure de convergence, en 2016 par ROBERT TATOIAN & LUTZ HAMEL [269] : l'indice de convergence *cix*, inspiré de la U-Matrix. Cet indice y ajoute une mesure appelée *Map Embedding Accuracy* permettant de comparer la distribution des neurones d'un SOM, après apprentissage, avec la distribution des données en entrée : en effet, YIN ET ALLISON [270] ont montré, en 1995, que la distribution des neurones d'un SOM suffisamment grand converge vers la distribution de probabilité des données d'entrée après un temps assez long.

Voisinage. Un SOM est un *réseau* de neurones, ou nœuds, et ceux-ci sont connectés entre eux. L'architecture SOM admet plusieurs implémentations du voisinage d'un nœud $r_{i,j}$. Il est notamment courant que les nœuds soient connectés selon le voisinage de VON NEUMANN (**Fig. 6.3**) ou de MOORE (**Fig. 6.4**). Chaque type de voisinage aura pour conséquence une modification de la façon dont la ressemblance entre une BMU et un vecteur d'entrée sera propagée aux nœuds voisins (cf. **Eq. 6.3**). D'autres types de voisinage peuvent également être considérés changeant drastiquement la topologie de la carte, notamment celle supprimant les effets de bords comme dans le cas des grilles utilisant un tore et faisant se connecter des nœuds situés aux opposés de la carte.

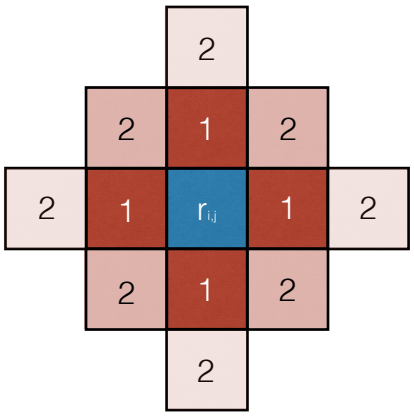


FIGURE 6.3 – VOISINAGE SELON VON NEUMANN — Voisinage d'un nœud $r_{i,j}$ selon le voisinage défini par Von Neumann, correspondant à une distance de Manhattan de portée n. Ici, $n = 2$.

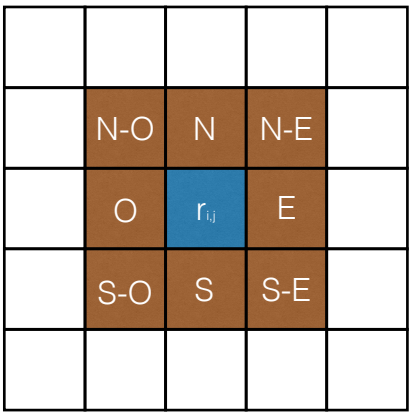


FIGURE 6.4 – VOISINAGE SELON MOORE — Voisinage d'un nœud $r_{i,j}$ selon le voisinage de Moore, correspondant à une distance de Chebychev de portée n. Ici, $n = 1$.

6.1.2 Discussion

Parmi les très nombreux algorithmes d'apprentissage existant, nous avons choisi d'utiliser une carte auto-organisatrice (ou SOM), algorithme formalisé et développé depuis longtemps et ayant été intensivement utilisé, adapté, enrichi et étudié par toute la communauté de l'apprentissage machine. Etant donné les contraintes de notre problème que nous avons exposées plus haut, cet algorithme offre un équilibre entre puissance et malléabilité. Cependant, tel quel, un SOM ne peut directement résoudre notre problème de fusion et d'inférence de données multimodales.

Premièrement, la plupart des utilisations faite des SOM sont effectuées hors-ligne et avec une matrice de donnée **P** entièrement disponible. Or, dans le cadre du modèle HTM, aucune donnée n'est connue préalablement à l'apprentissage, contraignant donc une utilisation en ligne. De plus, la matrice des données issues des capteurs visuels et audio (et traitée préalablement par des experts d'identification) est acquise vecteur par vecteur, au cours de l'exploration de l'environnement. Nous avons donc dû apporter quelques modifications permettant d'adapter le SOM aux caractéristiques de notre utilisation *en ligne*.

Une autre limitation apparait alors ici. Intégrant donc la capacité de l'utiliser en ligne, une carte auto-organisatrice pourrait nous permettre d'apprendre la relation entre la modalité audio et la modalité visuelle, selon un apprentissage en ligne et un paradigme de type non-supervisé. Cependant, un des buts du modèle HTM est d'être capable, à partir à cet apprentissage, d'*inférer* une modalité si celle-ci est manquante, par exemple lorsqu'un objet audiovisuel est situé derrière le robot donc inaccessible pour les capteurs visuels dont il est doté. Cette inférence permettra au système de toujours avoir accès à la représentation multimodale des événements apparaissant dans l'environnement et de pouvoir réagir avec pertinence à ceux-ci. Une autre modification a donc été nécessaire pour adapter le SOM traditionnel et le rendre capable d'effectuer un tel processus d'inférence. L'ensemble des modifications apportées au SOM traditionnel ont été formalisées par la *Multimodal-Self Organizing*

Map (M-SOM), décrit en détail à la section suivante.

6.2 Le Multimodal Self-Organizing Map

Le *Multimodal Self-Organizing Map* (M-SOM) est une adaptation d'une carte auto-adaptatrice (ou SOM), algorithme décrit à la section précédente. Les cartes de Kohonen sont de très puissants outils pour représenter de façon compacte des données de haute dimension. Cependant, l'algorithme classique utilisé pour le SOM présente un défaut majeur pour les besoins du modèle HTM : il n'est pas utilisable lorsque des données sont manquantes.

6.2.1 Pourquoi le SOM traditionnel n'est-il pas entièrement adapté ?

L'algorithme SOM, tel quel, ne permet pas de traiter le problème des données manquantes et ce, pour deux raisons :

1. les données d'entrées doivent toujours avoir la même dimension puisque :
   a. les vecteurs de poids des nœuds ont été initialisés avec une taille précise ;
   b. il n'est pas possible de calculer une distance entre deux vecteurs de dimensions différentes.
2. même en gardant un vecteur d'entrée de même dimension mais en mettant les composantes correspondantes à la modalité manquante à une valeur de 0 (ce qui serait acceptable puisque les probabilités d'appartenance à une classe étant nulle si la modalité est absente), cela ne résoudrait pas le problème car :
   a. le SOM considérerait cette entrée comme une autre ;
   b. le SOM coderait ainsi une distribution de probabilités dans laquelle toute une partie des composantes seraient nulles, correspondant à une classe audio ou visuelle inexistante.

Il n'est ainsi pas possible de retrouver une modalité manquante à partir d'une carte auto-adaptative classique. C'est pourquoi nous avons revisité l'approche du SOM pour l'adapter à cette contrainte tout en gardant ce qui en fait un outil extrêmement bien adapté à notre problématique.

Ainsi, nous introduisons le *Multimodal Self-Organizing Map* en tant qu'un enrichissement du SOM traditionnel.

6.2.2 Formalisation du M-SOM

Pour commencer, le M-SOM, utilise **deux vecteurs de poids** par nœuds r_{ij} : $\mathbf{w}^a_{ij}$ et $\mathbf{w}^v_{ij}$. Un vecteur de poids est dédié à la modalité audio, de dimension égale au vecteur de probabilités émis par l'AuditoryIdentityKS ; un second vecteur de poids est, lui, dédié à la modalité visuelle, de dimension égale au vecteur de

probabilités émis par la VisualIdentityKS. Cette architecture nouvelle peut être comprise comme la superposition de deux SOM traditionnels :

1. un premier dédié à la modalité audio avec ses propres nœuds r_{ij}^a
2. un second dédié à la modalité visuelle avec ses propres nœuds r_{ij}^v.

Nous définissons ici la notion de réseau global et de ses sous-réseaux constitutifs :

Définition 14. *Le* RÉSEAU GLOBAL *est la fusion d'un ensemble de sous-réseaux. Ce réseau code l'ensemble de l'information multimodale par représentation de la jonction des différents sous-réseaux qui le constituent.*

Définition 15. *Un* SOUS-RÉSEAU *est une carte auto-organisatrice dédiée à une modalité et une seule. Cette carte ne code donc qu'un seul type d'information.*

Les sous-réseaux seront ensuites fusionnés selon la définition suivante :

Définition 16. *Une catégorie audiovisuelle est dite « codée » par le M-SOM si et seulement si :* $r_{ij}^a = r_{ij}^v = r_{ij}$.

Autrement dit, une des modalités va servir de référence lors de la phase apprentissage et imposera la BMU de l'autre modalité. Cette particularité de l'apprentissage du M-SOM est une partie du paradigme global d'apprentissage auto-supervisé / par renforcement en cela qu'une modalité va diriger l'apprentissage de l'autre modalité afin que les deux puissent, *in fine*, de façon indépendante, être capables de retrouver la connaissance globale. Afin d'accomplir cela, une seule BMU sera déterminée et c'est celle-ci qui sera utilisée pour l'apprentissage de la seconde modalité. Ainsi, un même nœud sera capable de coder deux informations différentes si bien que lorsqu'une des deux est manquante, il sera possible de la retrouver. Selon ce nouveau paradigme, il est nécessaire de redéfinir la manière dont les vecteurs de poids sont appris. Les sections suivantes décrivent les deux cas ayant une influence sur la phase d'apprentissage et auxquels le module MFI sera confronté — et donc le M-SOM :

- si le robot a accès aux informations visuelles et auditives,
- si le robot n'a qu'un accès limité aux informations, c'est-à-dire si une des deux modalités n'est pas disponible.

6.2.2.1 Si toutes les modalités sont disponibles

Un événement $\ddot{\psi}_k[t]$ audiovisuel est émis par une source $\mathcal{S}_k$ et le robot fait face à cette source. Il a donc accès à toutes les modalités, se traduisant par des vecteurs $\mathbf{P}^a[t]$ et $\mathbf{P}^v[t]$ contenant des données issues des classifieurs concernant la même source $\mathcal{S}_k$. Dans ce cas, et dans ce cas seulement, le M-SOM sera capable d'apprendre le lien entre ces deux modalités.

Itération d'apprentissage. A partir de l'**Eq. 6.1**, nous définissons une BMU *audiovisuelle* r^{av}_{BMU} comme

$$\begin{aligned} r^{av}_{\text{BMU}} &= r_{IJ}[t], \text{ où} \\ (I,J) &= \arg\min_{i,j} \left(\|\mathbf{P}^a[t] - \mathbf{w}^a_{ij}\| \times \|\mathbf{P}^v[t] - \mathbf{w}^v_{ij}\| \right) \end{aligned} \tag{6.4}$$

La BMU est maintenant définie comme étant le nœud dont le vecteur de poids associé

$$\mathbf{w}^{av}_{\text{BMU}} = ({\mathbf{w}^a_{\text{BMU}}}^T, {\mathbf{w}^v_{\text{BMU}}}^T)^T$$

est composé des vecteurs audio et visuel les plus similaires, au sens de la distance euclidienne, aux vecteurs d'entrée $\mathbf{P}^a[t]$ *et* $\mathbf{P}^v[t]$ respectivement. Une fois la BMU déterminée, le reste de l'algorithme d'apprentissage reste le même (cf. **Eq. 6.2** & **Eq. 6.3**).

Estimation de la catégorie. D'une façon similaire au SOM traditionnel, la BMU d'un vecteur d'entrée inconnu nous permet justement de déterminer la catégorie, une fois le réseau appris. Dans le cas du M-SOM, il est nécessaire d'appliquer l'**Eq. 6.4** à la carte audio et à la carte visuelle afin d'obtenir les BMU correspondant au vecteur d'entrée audio et visuel, respectivement. Cependant, il est possible, et même probable, que ces neurones gagnants soient différents. En effet, un label audio peut par exemple être associé à plusieurs labels visuels (*male speech* et *female speech* par exemple) entraînant une non correspondance entre r^a_{BMU} et r^v_{BMU}. Il est donc nécessaire d'étudier la *BMU conjointe* des deux « sous-réseaux ». C'est ce qui se produit lors du calcul du neurone gagnant audiovisuel grâce à l'**Eq. 6.4**, prenant en compte les contributions des deux sous-réseaux à la distance euclidienne calculée. Cette nouvelle r^{av}_{BMU} permet donc d'estimer la catégorie audiovisuelle $\widehat{\mathcal{C}}^{(\text{all})}[t]$ [4].

6.2.2.2 Si une modalité est manquante

Un événement unimodal $\dot{\psi}_k$ est émis par une source $\mathcal{S}_k$ ou un événement $\ddot{\psi}_k$ est émis par une source $\mathcal{S}_k$ mais le robot n'a pas accès aux deux modalités (il ne fait pas face à la source, un obstacle empêche d'avoir accès aux données visuelles ou l'objet n'émet pas de son). Dans une telle situation, il est impossible de faire apprendre le M-SOM. Au lieu de cela, il est utilisé pour *inférer* la modalité manquante. Considérons le cas, par exemple, où il n'y a pas de modalité visuelle :

1. l'audio seul est utilisé ($\mathbf{P}^a[t]$) afin de déterminer la r^a_{BMU} dans le réseau audio, dont le vecteur de poids associé $\mathbf{w}^a_{\text{BMU}}$ peut être utilisé pour déterminer la catégorie audio $\widehat{\mathcal{C}}^a_A[t]$, with $A = \arg\max_k w^a_k$;
2. la BMU visuelle est directement déterminée par la BMU audio précédemment trouvée telle que $r^v_{\text{BMU}} = r^a_{\text{BMU}}$. *C'est précisément ici que l'apprentissage antérieur du lien entre les modalités est exploité* ;

4. où $^{(\text{all})}$ indique que les deux modalités sont disponibles.

3. enfin, le vecteur de poids $\mathbf{w}^v_{\text{BMU}}$ associé à la BMU visuelle r^v_{BMU} est utilisé pour déterminer la catégorie visuelle $\widehat{\mathcal{C}}^v_V[t]$, avec $V = \arg\max_l w^v_l$.

Le M-SOM est ainsi capable de fournir une estimation de la catégorie audiovisuelle $\widehat{\mathcal{C}}^{(\text{miss})}[t] = (\widehat{\mathcal{C}}^a_A[t], \widehat{\mathcal{C}}^v_V[t])$ [5], même lorsqu'une modalité est manquante. L'approche inverse lorsque la modalité audio est manquante est tout aussi valable.

6.2.3 Paramètres du M-SOM

Cette section détaille les paramètres à déterminer lors de l'utilisation d'un algorithme de type SOM.

6.2.3.1 Initialisation des vecteurs de poids

Généralement, il est possible d'initialiser les vecteurs de poids au démarrage de l'apprentissage selon deux méthodes : une méthode aléatoire et une méthode incluant une analyse préalable de la matrice de données à apprendre. Mais dans notre cas, nous n'avons pas accès aux données avant l'apprentissage puisque le système n'y a accès qu'au moment de l'exploration et de façon progressive. Cependant, bien que ne connaissant pas la façon dont la matrice des données perçues par le robot sera ordonnée, nous savons que ces données seront issues des experts de classification, experts dont nous connaissons le comportement. En effet, chaque KS d'identification utilisée par le logiciel TWO!EARS, KS présentées à la **Sec. 3.2.3**, envoient une probabilité que la trame audio au temps t appartienne à la catégorie pour laquelle la KS a été apprise. Lorsque nous rassemblons les probabilités de chaque KS, nous obtenons le vecteur $\mathbf{P}[t]$, comme décrit précédemment. Par conséquent, il est également possible de définir une catégorie audiovisuelle comme un vecteur de $N_a + N_v$ probabilités dans lequel seulement deux composantes (une pour la partie audio et une pour la partie visuelle) seront significativement supérieures aux autres. Sachant cela, il est possible de considérer l'introduction de cette connaissance au préalable *via* l'initialisation des vecteurs de poids : chaque vecteur de poids est initialisé selon un vecteur aléatoire dont toutes les composantes sauf une sont inférieures à 0.5, la composante restante, choisie aléatoirement, étant fixée à 1 (probabilité maximum). Cette initialisation « biaisée » doit en revanche prendre en compte l'architecture particulière du M-SOM : celui-ci possède deux vecteurs de poids par nœuds et non un seul. Il n'est alors pas possible d'initialiser indépendamment les vecteurs de poids des nœuds des deux sous-réseaux puisque cela introduirait un biais dans l'appariement entre une classe audio et une classe visuelle. Ce mauvais appariement, représenté par la conjonction des nœuds des deux sous-réseaux complexifierait la tâche d'apprentissage car la probabilité que la composante maximale introduite dans un nœud du sous-réseau audio corresponde effectivement à la bonne classe visuelle (celle qui sera observée lors de l'exploration) est beaucoup trop faible. Une solution serait alors d'effectuer l'initialisation biaisée seulement sur un seul des deux sous-réseaux et d'initialiser l'autre sous-réseau soit avec des valeurs aléatoires, soit avec des composantes toutes égales, afin de ne laisser qu'une modalité « diriger » l'apprentissage.

5. où $^{(\text{miss})}$ indique qu'une modalité est manquante

Une initialisation prenant en compte la connaissance sur le comportement des experts d'identification permettrait également de résoudre un second problème. Lors d'une initialisation aléatoire, les vecteurs de poids contiennent souvent plusieurs « pics », pics similaires à ceux présents dans les données à traiter. Or, au cours des toutes premières étapes d'apprentissage du M-SOM, un nœud gagnant est déterminé, la BMU, et sa ressemblance au vecteur d'entrée est augmentée, ayant pour effet de favoriser le pic correspondant au vecteur considéré et d'estomper tous les autres. Conséquemment, après apprentissage de cet exemple, une petite zone de la carte possède une répartition de ses poids beaucoup plus similaire aux données en entrée que le reste du réseau. Cela a pour conséquence d'attirer tout l'apprentissage sur cette zone et de restreindre grandement la propagation de l'apprentissage et de condenser toutes les catégories apprises dans cette petite zone. Initialiser selon la méthode que nous avons appliquée permettrait d'assurer, particulièrement au début de l'apprentissage — phase critique — que n'importe quelle zone du réseau peut être un bon candidat pour la BMU des vecteurs d'entrée.

Après de nombreux tests sur des scénarios allant du plus simple (trois sources, deux catégories audiovisuelles différentes, cas unisource) au plus complexe (dix sources, sept catégories audiovisuelles différentes, cinq sources simultanées), aucune différence majeure concernant la rapidité de convergence ou la propagation de l'apprentissage n'a été constatée entre une initialisation aléatoire et une initialisation prenant en compte le comportement connu des experts d'identification. Nous avons donc opté pour une initialisation aléatoire afin d'introduire un biais minimum dans l'auto-organisation future du réseau.

6.2.3.2 Nombre d'itérations

Les SOM traditionnels fonctionnent généralement avec des matrices importantes de données recueillies avant la phase d'apprentissage, dans le but de les catégoriser. D'autre part, le fait que les SOM traditionnels fonctionnent hors-ligne permet de réaliser une étape d'optimisation des paramètres d'apprentissage, en particulier le nombre d'itérations nécessaire pour que le réseau converge. Cette convergence est estimée par analyse des groupes créés lors de l'apprentissage : un SOM converge lorsque les frontières des groupes ne change plus d'une itération d'apprentissage à une autre.

Or le contexte dans lequel le modèle HTM fonctionne est différent en cela que les données perçues doivent être analysées le plus rapidement possible afin de faire naître une réaction du robot, *via* des mouvements de tête. Il n'est donc pas possible :

1. d'attendre d'avoir un nombre important de données (et comment estimer quand ce nombre serait atteint ?) ;
2. de juger du nombre d'itérations nécessaire pour que le réseau converge puisque le nombre de catégories audiovisuelles que le réseau va créer est toujours susceptible de changer ;
3. de fixer un nombre d'itérations trop important puisque plus ce nombre est grand, plus le temps de calcul est grand.

Du fait de ces contraintes, le fonctionnement du M-SOM a également dû être adapté par rapport au SOM traditionnel.

1. Le M-SOM va n'apprendre que le dernier vecteur audiovisuel perçu. Les données précédemment collectées ne sont donc pas prises en compte et l'apprentissage reprend là où il s'était arrêté.
2. Concernant le nombre d'itérations nécessaires pour que le réseau apprenne correctement, ce nombre a été fixé à $N_{it} = 10$ (quelques précisions importantes sur ce paramètre sont données plus bas). Comparativement avec la littérature, ce chiffre est très bas : généralement, le nombre d'itérations varie entre une centaine et plusieurs milliers. L'étude de l'impact du nombre d'itérations sur la qualité de l'apprentissage ainsi que sur la qualité du comportement du robot (la qualité de l'apprentissage ayant un impact sur ses mouvements de tête) sera effectuée à la **Sec. 6.5**. D'autre part, un faible nombre d'itérations permet également un gain de temps computationnel lors de l'étape d'apprentissage de la dernière trame audiovisuelle.

Concernant le dernier point, nous souhaiterions préciser la façon dont les itérations d'apprentissage ont été gérées. L'algorithme SOM traditionnel effectue N_{it} itérations d'apprentissage pour toutes les données à apprendre. Etant données les caractéristiques de notre problème, nous avons apporté une modification substantielle jouant également un rôle dans la gestion des éventuelles erreurs de classification des experts d'identification TWO!EARS, particulièrement ceux dédiés à l'audio. Ces experts d'identification audio nécessitent généralement un certain temps (assez court) avant de converger vers la bonne catégorisation. Ainsi, les toutes premières classifications d'une source audio auront plus de chances d'être fausses. Nous avons intégré cette dynamique par une gestion différente du caractère itératif de l'apprentissage du M-SOM. Deux modifications ont ainsi été apportées.

Premièrement, nous avons défini une valeur d'itération $n_{it}^{o_j}$ par objet créé par le module MFI. Ainsi, l'apprentissage des données concernant d'un objet sera indépendant de l'apprentissage d'un autre. D'autre part, dans l'algorithme traditionnel, les premières itérations sont celles pour lesquelles la largeur de la gaussienne et le taux d'apprentissage sont maximum. Nous avons effectué l'inverse : les premiers résultats de classification des experts d'identification étant ceux pour lesquels le taux d'erreur risque d'être le plus élevé, nous avons fait varier l'itération de façon inverse. Nous définissons ainsi la valeur de l'itération $n_{it}^{o_j}$ de l'objet o_j selon :

$$n_{it}^{o_j}[t] = t_i + (t - t_i) \tag{6.5}$$

La valeur de l'itération n_{it} de l'apprentissage d'un vecteur appartenant à l'objet o_j est définie comme :

$$n_{it}[t] = \max\left((N_{it} - n_{it}^{o_j}[t]) + 1, 1\right) \tag{6.6}$$

Ces deux modifications de l'évolution du nombre d'itérations auront un effet sur la correction des erreurs de classification, sur la rapidité de l'exécution de l'étape d'apprentissage puisque que pour un nouveau vecteur à apprendre, une seule itération d'apprentissage sera effectuée.

6.2.3.3 Paramètres d'apprentissage

L'algorithme d'apprentissage du M-SOM est similaire à celui du SOM et nécessite la détermination de quatre paramètres :

1. nombre de neurones du réseau : nous avons choisi de créer un réseau de taille $(\lceil\sqrt{N_a \times N_v}\rceil \times \lceil\sqrt{N_a \times N_v}\rceil)$ afin de pouvoir avoir au minimum un neurone codant pour chaque combinaison de catégories audio et visuelles possible ;
2. fonction de voisinnage pour la connexion des neurones entre eux : comme introduit à la **Sec. 6.1.1**, plusieurs conceptions de la notion de voisinnage existent dans le cadre d'un algorithme de type SOM. Nous avons choisi le voisinnage de *Moore*, ou distance de Manhattan de portée égale à 2 ;
3. taux d'apprentissage α : nous avons choisi $\alpha \in [0.9, 0.02]$ (valeurs décroissantes en fonction de l'itération d'apprentissage), inspiré de [208] ;
4. largeur de la gaussienne σ matérialisant l'ampleur de la propagation de l'apprentissage : nous avons choisi $\sigma \in [3, 1]$ (valeurs décroissantes en fonction de l'itération d'apprentissage), permettant de propager à trois voisins l'apprentissage de la BMU.

6.2.4 Convergence de l'apprentissage

Une des étapes majeures, et délicates, des algorithmes d'apprentissage non-supervisés est la définition de sa *convergence*. En effet, là où les algorithmes supervisés admettent une limite imposée par l'expérimentateur au-delà de laquelle le réseau est considéré comme *appris*, en fonction d'un état final que le réseau doit atteindre, les algorithmes non-supervisés, en revanche, ne possèdent pas de représentation finale vers laquelle ils doivent tendre. Sachant que les mouvements de tête sont générés en réponse du besoin d'apprentissage du réseau et que le but du modèle HTM est, *in fine*, d'inhiber ces mouvements de tête, il est indispensable de définir une limite au-delà de laquelle le réseau est considéré comme appris, ne requérrant ainsi plus de mouvements de tête.

Cependant, l'apparition d'événements dans l'environnement en cours d'exploration étant totalement imprévisible, stopper intégralement l'apprentissage du réseau pourrait conduire à l'empêcher de créer de nouvelles catégories audiovisuelles dans le cas où elles apparaîtraient. Ainsi, plutôt que de définir un seuil à partir duquel le réseau *entier* arrête d'apprendre, nous avons défini un critère permettant au module MFI de juger **catégorie par catégorie** si l'apprentissage local doit être arrêté ou poursuivi : le critère q.

Le critère q permet de déterminer si une catégorie audiovisuelle $c^{(i)}(a_j, v_j)$ nécessite une poursuite de l'apprentissage du lien entre les composantes audio et visuelles ou si cet apprentissage est désormais inutile. Soit le symbole de Kronecker $\delta_{ij}^{(k)}[t]$ défini comme :

$$\delta_{ij}^{(k)}[t] = \begin{cases} 1 \text{ si } \widehat{\mathcal{C}}^{(k)}[t] = (\mathcal{C}_i^a[t], \mathcal{C}_j^v[t]), \\ 0 \text{ sinon.} \end{cases} \tag{6.7}$$

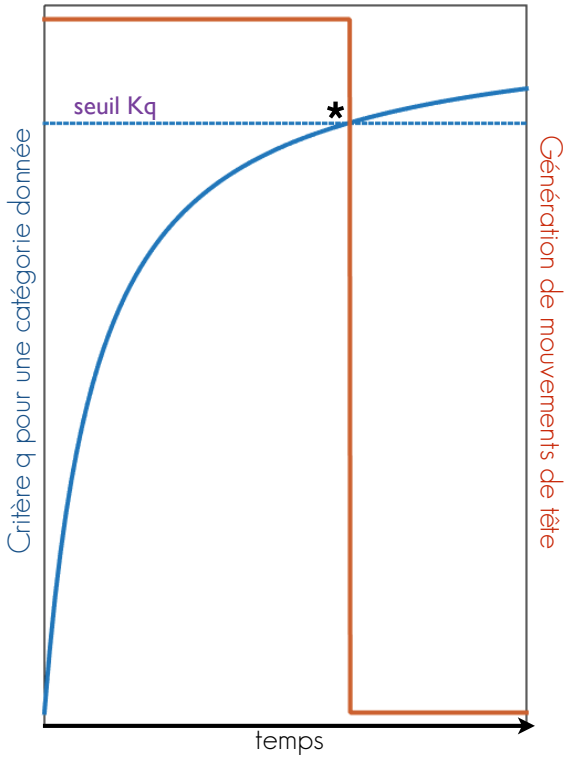


FIGURE 6.5 – CRITÈRE q ET SEUIL K_q — Illustration schématique de l'impact du critère q sur la génération de mouvements de tête, pour une catégorie donnée : (*bleu*) critère q, (*rouge*) nombre de mouvements de tête, ($*$) point où $q = K_q$. C'est à partir de ce point que le module MFI juge qu'il peut faire confiance en ses connaissances sur cette catégorie audiovisuelle et ainsi ne plus requérir de mouvements de tête.

où $k = \{\text{all}, \text{miss}\}$ indique si la catégorie a été obtenue en présence de toutes les modalités ou non. Nous définissons le *ratio d'inférence* $q(\widehat{\mathcal{C}}^{a,v})[t]$ de la catégorie audiovisuelle $(\mathcal{C}_i^a, \mathcal{C}_j^v)$ au temps t comme :

$$q(\widehat{\mathcal{C}}^{a,v})[t] = \frac{\sum_1^t \delta_{ij}^{(\text{miss})}[t-1]\, \delta_{ij}^{(\text{all})}[t]}{\sum_1^t \delta_{ij}^{(\text{miss})}[t]} \tag{6.8}$$

Ainsi, $q(\widehat{\mathcal{C}}^{a,v})$ est le ratio entre le nombre d'inférences confirmées et le nombre total d'inférences que le M-SOM a effectuées pour cette catégorie audiovisuelle.

Définition 17. *Inférence confirmée : Catégorie audiovisuelle au temps $t-1$ inférée par le M-SOM qui a été confirmée, après mouvement de tête et accès à toutes les modalités, au temps t.*

La **Fig. 6.5** illustre le comportement issu de la définition de ce critère q. Ce critère permettra donc d'avoir une idée sur la capacité du module MFI, au temps t, d'inférer correctement une catégorie audiovisuelle. Or, ce critère considéré seul ne permet pas la prise de décision sur l'arrêt de l'apprentissage ou sa poursuite. C'est pourquoi nous le comparons à un seuil déterminé $K_q \in \mathbb{R}^+ = [0, 1]$ donnant au robot la capacité

de modifier la façon dont il va explorer l'environnement en cours. Si la valeur du critère q, pour une catégorie audiovisuelle donnée, dépasse ce seuil, le module MFI considèrera que le système a suffisamment appris cette catégorie pour inhiber les mouvements de tête. A l'opposé, tant que ce critère q est en-dessous de cette valeur, le module MFI va générer des mouvements de tête, jusqu'à ce que le système soit suffisamment performant.

Plus le seuil K_q est élevé, plus le nombre d'inférences confirmées nécessaires pour que le critère q le dépasse est grand. Cela implique donc un plus grand nombre de mouvements de tête et, potentiellement, une plus grande source de distraction pour le robot. Par exemple, un seuil fixé à $K_q = 0.8$ signifie que, pour une catégorie audiovisuelle donnée, 80% des inférences la concernant doivent être confirmées avant que le module MFI n'arrête l'apprentissage de cette catégorie.

Ce critère, plutôt qu'un paramètre restreignant le fonctionnement du système, peut être interprété comme une façon de lui donner la possibilité d'adapter son comportement en fonction de la tâche qu'il a à effectuer. Dans un scénario de type S&R, l'exploration est importante mais pas primordiale. Ainsi, un seuil bas permettra au robot de générer quelques mouvements de tête mais de rapidement faire confiance en son inférence, quitte à n'atteindre qu'un apprentissage incomplet et à effectuer quelques erreurs de catégorisation. Cependant, il pourra se consacrer plus pleinement à l'objet prioritaire nécessitant son attention (une victime par exemple). A l'opposé, dans des scénarios où le robot n'est pas contraint en terme de temps ou de priorité de tâche à effectuer, le seuil pourra être très haut afin qu'il apprenne le mieux possible son environnement.

Enfin, nous avons présenté à la **Sec. 6.1.1** deux mesures de la convergence d'un réseau de type SOM : la U-Matrix traditionnelle et l'index de convergence introduit par Tatoian & Hamel [269]. L'utilisation de ces critères ne serait pas pertinente ici puisque nous ne cherchons pas à faire totalement converger le M-SOM : une zone doit toujours être disponible pour permettre l'inclusion de nouvelles catégories audiovisuelles. De plus, lorsque nous calculons la U-Matrix du M-SOM comme illustré par la partie gauche de la **Fig. 6.6**, nous observons deux zones : une correspondant à l'espace des données apprises (jaune), une autre correspondant aux nœuds n'ayant pas été touchés par l'apprentissage (bleu). Cette variation importante induit un aplanissement de la représentation et nous empêche de distinguer clairement la topologie de la zone apprise.

Cependant, nous pouvons effectuer une mesure locale de la convergence du M-SOM en ne prenant que la zone du réseau ayant subi un apprentissage (partie jaune) Cette mesure locale a été faite selon plusieurs étapes :

1. calcul la U-Matrix du réseau M-SOM entier (combinaison des deux sous-réseaux) ;
2. normalisation ;
3. homogénéisation de toutes les valeurs de distances supérieures à un seuil (dépendant des données).

Cette modification, illustrée par la partie droite de la **Fig. 6.6**, nous permet de mieux visualiser la façon dont le M-SOM s'auto-organise en mettant mieux en valeur la topologie du réseau.

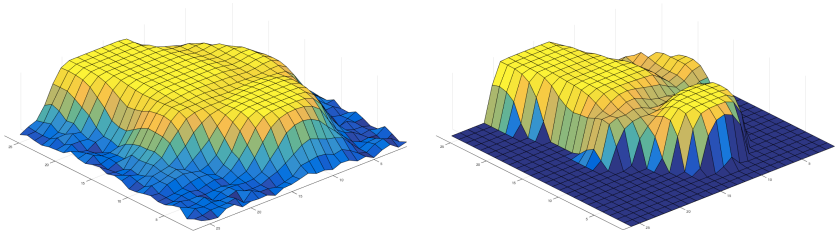


FIGURE 6.6 – U-MATRIX CLASSIQUE ET MODIFIÉE — (*gauche*) calcul de la U-Matrix classique, (*droite*) calcul de la U-Matrix incluant un seuillage. Par souci de clarté, l'opposé des distances a été représentée : (*jaune*) distances faibles, (*bleu*) distances élevées. Ce seuillage permet de mettre en exergue la topologie du réseau, c'est-à-dire les différentes zones codant une même catégorie audiovisuelle.

6.2.5 Discussion

Cette section a décrit en détail l'architecture du *Multimodal Self-Organizing Map*, algorithme d'apprentissage inspiré d'une carte auto-organisatrice (SOM) dont le fonctionnement a été décrit à la **Sec. 6.1**. Tirant partie de leur puissance et de la littérature abondante permettant d'avoir une formalisation solide — ce qui n'est pas le cas pour tous les algorithmes d'apprentissage machine — nous avons proposé une adaptation du SOM traditionnel permettant d'y inclure la multimodalité. La multimodalité a été introduite en divisant le SOM en deux sous-réseaux, l'un dédié à l'apprentissage des données audio, l'autre à celui des données visuelles. La façon dont nous avons « orienté » l'apprentissage, c'est-à-dire en imposant qu'une modalité dirige l'apprentissage des deux modalités, est un des éléments du paradigme plus global dans lequel le module module MFI se situe : celui d'apprentissage auto-supervisé / par renforcement. En effet, l'algorithme que nous avons développé permet d'apprendre le lien qui existe entre deux modalités. Ce faisant, lorsqu'une d'entre elles est manquante, il nous est possible de la retrouver, sur la base unique de l'expérience passée du système. Reprenant l'exemple du pianiste débutant qui effectue une association sensorimotrice entre le mouvement de ses doigts et la perception auditive des notes jouées, exemple classique de l'apprentissage par renforcement et tiré de la **Sec. 2.4**, le M-SOM consisterait en la création d'un sous-réseau dédié à la perception motrice et d'un second sous-réseau dédié à l'écoute des notes jouées. Ainsi, une fois l'apprentissage suffisamment avancé, si jamais la modalité audio est manquante, la simple sensation motrice suffirait à déterminer si les notes jouées sont les bonnes ou non.

Un autre intérêt de notre approche réside dans le fait de pouvoir l'étendre à d'autres modalités : le M-SOM n'est pas contraint à n'avoir que deux sous-réseaux. Des expériences préliminaires sur l'ajout de deux autres sous-réseaux dédiés à la localisation audio et visuelle, séparément, ont montré que le M-SOM est tout à fait capable de fournir une représentation plus complexe des objets multimodaux que le robot perçoit. Conceptuellement, le M-SOM serait en mesure de prendre en compte n'importe quelle source d'information participant à la définition d'un objet. Dans le cadre de cette thèse, seule l'inclusion des modalités audio et visuelles ont été testées et validées.

Précédemment, à la **Sec. 4.4.2**, nous avons introduit le fait que le modèle HTM soit capable d'effectuer la fusion des experts d'identification audio et visuels. En effet, le M-SOM, par construction, est un algorithme permettant de prendre une décision sur l'appartenance d'un vecteur d'entrée de haute dimension à une catégorie unique. La section suivante détaille donc cette étape de fusion, étape majeure du modèle HTM.

6.3 Fusion de classifieurs

Le module MFI, sur la base du *Multimodal Self-Organizing Map* décrit ci-dessus, effectue une étape de fusion de classifieurs intermodale. Cette fusion est formalisée par un système de type *Decision Support System* (DSS, cf. **Sec. 2.4.2**) permettant de prendre une décision binaire sur l'appartenance d'un vecteur de probabilités d'appartenance aux classes audio et visuelles donné par les experts d'identification. Les notations employées ici ont été introduites à la **Sec. 4.4.2**.

6.3.1 Fusion intramodale

La fusion intramodale consiste ici en la prise de décision, à partir du vecteur de probabilités émis par un expert d'identification, de la composante gagnante, c'est-à-dire celle qui sera le représentant de la catégorie unimodale audio ou visuelle. Autrement dit, parmi toutes les catégories audio que les experts d'identification audio sont capables de reconnaître, quelle est celle à laquelle l'objet appartient, sachant qu'il ne peut être caractérisé que par une seule catégorie par modalité ? Cette fusion est réalisée en deux étapes.

La première étape consiste en la façon dont les données sont organisées, objet par objet. Comme décrit à la **Sec. 4.2**, les définitions et propriétés d'un Objet dans le cadre du modèle HTM nous permettent d'effectuer une intégration temporelle des données perçues. Cette intégration est faite par la prise en compte de l'expérience passée du Robot concernant un objet audiovisuel donné. Ainsi, le vecteur $\mathbf{P}(o_j)[t]$ appartenant à l'objet o_j au temps t qui sera envoyé sera modifié selon l'**Eq. 4.7**. Cette première étape permet de pondérer le vecteur issu des experts d'identification au temps t en fonction de l'expérience du robot afin de diminuer l'impact d'erreurs momentanées de classification de ces experts.

La deuxième étape de fusion intramodale est réalisée par le M-SOM. Tout comme un SOM traditionnel, le M-SOM s'auto-organise au fur et à mesure de l'apprentissage. Cette organisation aboutira à la formation de zones distinctes constituées de plusieurs nœud codant la même information. La réduction de l'espace d'entrée en un espace de plus faible dimension est la deuxième étape de cette fusion intramodale de classifieurs en cela qu'à partir d'un vecteur de probabilités audio ou visuelles, chacun des sous-réseaux va lui faire correspondre une catégorie unique. Ainsi, cette étape de fusion consiste en la prise de décision, séparément, $d^a[t] \in [1, \ldots, N_a]$ sur le vecteur $\mathbf{P}^a[t]$, et $d^v[t] \in [1, \ldots, N_v]$ sur le vecteur $\mathbf{P}^v[t]$. Cette prise de décision est similaire à la détermination de la BMU^a et de la BMU^v respectivement.

Mais en quoi cette fusion diffère-t-elle d'une fusion effectuée en sortie directe des classifieurs par détermination de la composante maximum du vecteur $\mathbf{P}^a[t]$ et du vecteur $\mathbf{P}[t]$? En rien. A ce point, la fusion faite par chacun des sous-réseaux du M-SOM correspond à une décision prise en sortie directe des experts d'identification. En effet, ces deux sous-réseaux sont basés sur l'algorithme SOM réalisant une forme de quantification vectorielle aboutissant à la détermination d'une catégorie donné par sélection de la composante maximum du vecteur de poids de chaque neurone. Pour le moment donc, la fusion intramodale effectuée par le M-SOM est imparfaite puisqu'elle n'est pas en mesure de gérer les éventuelles erreurs de classification audio ou visuelle. En revanche, l'étape d'apprentissage de chaque sous-réseau est indispensable en cela que les neurones vont apprendre la distribution totale des vecteurs $\mathbf{P}^a[t]$ et $\mathbf{P}[t]$, incluant donc parfois les erreurs de classification des experts. Cette inclusion va justement nous permettre d'effectuer une fusion globale des classifieurs correcte, grâce à l'étape de fusion intermodale décrite ci-dessous.

6.3.2 Fusion intermodale

La fusion intermodale consiste quant à elle à déterminer la catégorie multimodale à laquelle un vecteur $\mathbf{P}(o_j)[t]$ appartient. Cette étape consiste en la mise en commun des contributions des sous-réseaux audio et visuel à la catégorisation de ce vecteur afin de calculer la décision $d^c[t]$. La fusion intermodale n'est réalisée que lorsque toutes les modalités sont présentes (si une modalité est manquante, le module MFI réalise alors une inférence et non une fusion). Dans le cas de données complètes, la détermination du neurone r^{av}_{BMU} par l'**Eq. 6.4** permet d'obtenir la catégorie audiovisuelle estimée par le M-SOM. Le fait d'effectuer une fusion à partir des distances audio et visuelles induit la propriété suivante :

Propriété 6. *La BMU^{av}, définie comme le maximum de la distance combinée des deux sous-réseaux au vecteur d'entrée* $\mathbf{P}[t]$*, n'est* pas forcément *égale à la concaténation de la BMU^a et de la BMU^v.*

La **Fig. 6.7** est la représentation des distances [6] calculées entre un vecteur de données observées appartenant à la catégorie audiovisuelle *male speech* et les vecteurs des poids du sous-réseau audio (en haut), du sous-réseau visuel (au centre) ainsi que la combinaison de ces deux matrices de distances. Les cercles noirs représentent ainsi respectivement les BMU^a, BMU^v et BMU^{av}. Cet exemple illustre la **Prop. 6** en cela que le nœud gagnant du sous-réseau audio est différent du nœud gagnant du sous-réseau visuel. La combinaison des deux inclut les contributions relatives des deux sous-réseaux à la ressemblance entre le vecteur d'entrée et chacun des sous-réseaux.

En quoi cette combinaison des sous-réseaux permet-elle de gérer d'éventuelles erreurs de classification audio ou visuelle ?

Nous rappelons que chaque expert d'identification est dédié à la détection et à la reconnaissance d'une catégorie audio ou visuelle donnée et que la probabilité émise par un expert est totalement indépendante des probabilités émises par les autres

6. A des fins de clarté, l'opposé des distances a été représenté : les zones les plus élevées correspondent aux nœuds ayant entre eux la distance la plus faible.

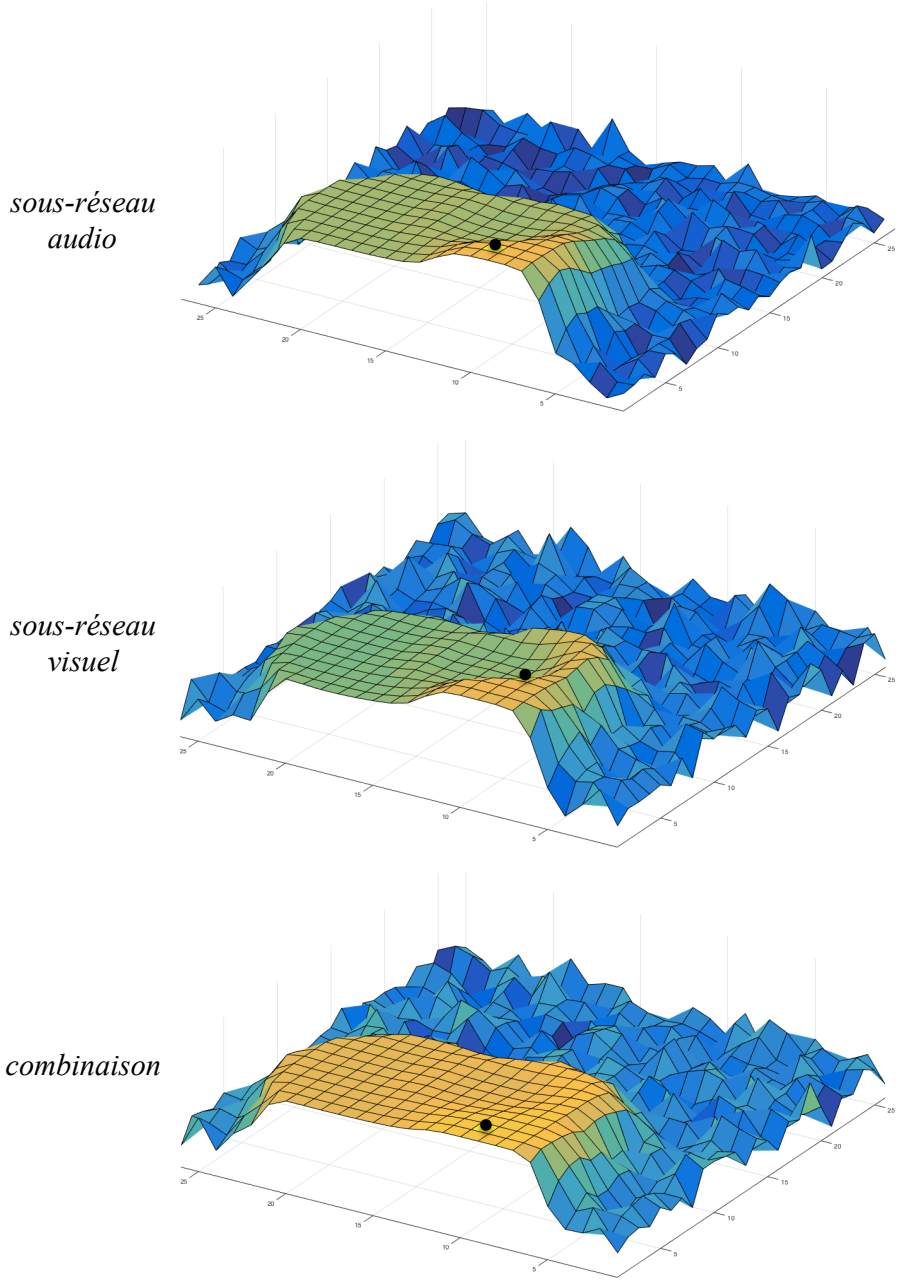


FIGURE 6.7 – MSOM, RÉSEAU GLOBAL ET SOUS-RÉSEAU — Distances euclidiennes normalisées entre un vecteur appartenant à la catégorie *male speech* et (*haut*) les nœuds du sous-réseau audio, (*centre*) les nœuds du sous-réseau visuel ; (*bas*) combinaison des deux matrices de distances obtenues ; (*cercle noir*) nœud gagnant, c'est-à-dire celui dont le vecteur de poids associé à la distance au vecteur d'entrée la plus faible. A des fins de clarté, l'opposé des distances a été représenté : les zones les plus élevées de la carte correspondent aux nœuds ayant la distance la plus faible.

experts, pour une trame donnée. Ainsi, une erreur de classification d'une trame au temps t consiste la plupart du temps en l'apparition d'un ou plusieurs pics erronés dans les vecteurs $\mathbf{P}^a[t]$ et $\mathbf{P}^v[t]$, pics surpassant parfois celui correspondant à la bonne catégorie (c'est-à-dire que la probabilité d'appartenance la plus élevée correspond à une mauvaise classe audio ou visuelle). De plus, nous rappelons que la

classification des experts est plus souvent correcte qu'erronée (cf. **Sec. 3.2.3**) et que le modèle HTM organise les données perçues selon la notion d'Objet (cf. **Sec. 4.2**). Ainsi, lors de l'apprentissage, les sous-réseaux auront plus souvent appris des distributions de probabilités (i) correspondant aux bonnes catégories (entre 70% et 90% théoriquement, pour les véritables experts d'identification, mais nous verrons lors de l'évaluation du modèle HTM en conditions réelles — donc utilisant ces experts — que ce taux chute aux alentours de 40%) et (ii) ayant inclus les éventuelles erreurs de classification. L'intérêt du critère de décision basée sur une distance et non directement sur la composante maximale permet donc justement de prendre en compte la distribution entière des vecteurs d'entrée et d'intégrer ces valeurs erronées dans la définition même d'une catégorie audiovisuelle, diminuant ainsi l'impact momentané qu'une erreur de classification peut avoir. En effet, ce n'est pas forcément la composante la plus forte qui l'emporte lors du calcul du minimum des distances mais bien l'ensemble du vecteur d'entrée.

La confiance que le module MFI porte en sa connaissance d'une catégorie audiovisuelle donnée est exprimée en fonction de la qualité de l'inférence d'une modalité manquante à partir d'une modalité présente, inférence permettant ainsi de reconstituer la catégorie audiovisuelle entière. Lorsque le module MFI ne fait pas confiance en son inférence, les mouvements de tête seront utilisés afin d'accéder aux informations visuelles manquantes. La section suivante décrit la façon dont le module MFI génère des ordres moteurs.

6.4 Ordre moteur

Comme expliqué à la **Sec. 6.2.2.2**, aucune phase d'apprentissage n'est lancée lorsqu'une modalité est manquante. En revanche, une tentative d'inférence est faite pour estimer la catégorie $\hat{\mathcal{C}}$ de l'objet considéré par le robot. Cependant, cette inférence peut se révéler fausse, particulièrement en début d'apprentissage. C'est pourquoi un critère de confiance en l'inférence effectuée a été développé : le critère $q(\widehat{\mathcal{C}}^{a,v})$, calculé pour chaque catégorie grâce à l'**Eq. 6.8** (cf. **Sec. 6.2.4**). En fonction de la confiance que le module MFI porte en son inférence d'une catégorie audiovisuelle, un ordre moteur peut être généré. Cette action motrice a pour but d'obtenir les données correspondant à la modalité manquante et une étape d'apprentissage par le M-SOM pourra conséquemment être lancée (cf. **Sec. 6.2.2.1**), toutes les modalités étant désormais présentes. De plus, la catégorie audiovisuelle observée (estimation *a posteriori*) pourra être comparée à l'inférence que le module MFI a effectuée avant le mouvement de tête (estimation *a priori*).

Le critère q sera mis à jour pour chaque source $\mathcal{S}_k$ émettant un son au temps t. A l'issue de cette mise à jour, il est possible, et fortement probable, que plusieurs sources requièrent un ordre moteur. Il est donc nécessaire de prendre une décision sur la source prioritaire.

Selon une approche similaire à celle utilisée pour le module DW, nous allons déterminer l'ordre moteur à générer par le module MFI grâce au modèle GPR. Chaque ordre moteur possible est une action motrice représentée par un canal d'information possédant sa propre activité. Soit $\mathbf{\Theta}_{\text{MFI}} \in \mathbb{R}^+ = [0, 359]$ l'ensemble des ordres moteurs possibles vers les n_s sources audiovisuelles émettant un son au temps t, tel que :

$$\mathbf{\Theta}_{\text{MFI}}[t] = [\theta(\mathcal{S}_1)[t], \ldots, \theta(\mathcal{S}_{n_s})[t]] \tag{6.9}$$

où chacun des angles est donné par l'expert de localisation audio. Selon l'approche considérant chaque ordre moteur comme un canal dédié, nous définissons l'activité $\tau_{\text{MFI}}[t]$ de la énième source $\mathcal{S}_n$ au temps t selon :

$$\tau_{\text{MFI}}(\mathcal{S}_n)[t] = \frac{q(\widehat{\mathcal{C}}^{a,v})[t]}{K_q} \tag{6.10}$$

avec $\widehat{\mathcal{C}}^{a,v}$ la catégorie de la source $\mathcal{S}_n$ et $q(\widehat{\mathcal{C}}^{a,v})[t]$ défini à l'**Eq. 6.8**. Plus l'activité τ_{MFI} d'une action motrice est forte, plus le système est confiant en la connaissance qu'il a de la catégorie audiovisuelle à laquelle cette source appartient. Lorsque l'activité dépasse la valeur seuil de 1, c'est-à-dire lorsque $q \geq K_q$, le module MFI inhibe la commande motrice vers cette source. Le canal gagnant est celui ayant l'activité la plus faible, de façon similaire au modèle GPR. Soit $\mathcal{S}_{min}$ la source ayant le taux d'activité le plus faible, l'ordre moteur d'angle $\theta_m[t]$ sera donc déterminé selon :

$$\theta_{\text{MFI}}[t] = \mathbf{\Theta}_{\text{MFI}}(\mathcal{S}_{min}) \tag{6.11}$$

Enfin, et de façon identique au module DW, une forme de persistance a été intégrée, similaire à celle induite par la boucle de rétrocontrôle positif dans laquelle le thalamus joue un rôle important [34, 52, 51]. Ainsi, lorsqu'une action motrice est générée, le canal correspondant aura tendance à s'auto-promouvoir durant un certain temps t_p afin d'éviter la génération éventuelle de mouvements contradictoires dans des temps très courts. Soit ainsi $n = t - t_i(\mathcal{S}_k)$ le temps écoulé depuis la première focalisation et $\delta^{(k)}(n)$ le symbole de Kronecker lié au k-ième canal à un temps :

$$\delta^{(k)}(n) = \begin{cases} -1 \text{ si } n < t_p, \\ 1 \text{ sinon.} \end{cases} \tag{6.12}$$

où $t_p = 10$ (mais d'autres valeurs sont possibles, comme la **Sec. 6.5.3** le montrera). L'activité du canal de la source $\mathcal{S}_k[t]$ sera donc désormais exprimée selon :

$$\tau_{\text{MFI}}(\mathcal{S}_k)[t] = \frac{q(\widehat{\mathcal{C}}^{a,v}(\mathcal{S}_k))[t]}{K_q} \times \delta^{(k)}(t - t_i) \tag{6.13}$$

Ainsi, dès lors que l'activité d'un canal correspondant à la source $\mathcal{S}_k$ a entraîné sa sélection pour générer un ordre moteur, c'est-à-dire au temps $t_i(\mathcal{S}_k)$, l'utilisation de l'**Eq. 6.12** permet d'introduire une boucle de rétroaction positive pour une certaine durée en baissant l'activité de ce canal. L'activité du canal soumis à cette boucle de rétroaction sera en effet systématiquement inférieure à 0, puisque lorsque $q(\mathcal{C}^{a,v})[t]$ est supérieur à K_q, l'activité sera ainsi supérieure à 1 et tout mouvement de tête sera inhibé.

Cette persistance aura ainsi un effet direct sur le nombre de mouvements de tête générés par le module MFI. Cette persistance permet surtout d'éviter les situations dans lesquelles deux objets appartenant à deux catégories différentes apparaissent à peu près au même moment et auront ainsi éventuellement une évolution de leur performance $q(\mathcal{C}^{a,v})$ similaire. Prenons le cas où lorsque deux canaux ont la même activité, la canal gagnant sera celui correspondant à l'objet ayant apparu le plus récemment, de façon similaire au module DW. Ainsi, comme illustré à la **Fig. 6.8**, la sélection de l'ordre moteur sur la base des activités de ceux deux canaux sera sans cesse partagée entre eux deux et occasionnera ainsi de très nombreux mouvements de tête, comportement que nous cherchons justement à éviter. Incluant le processus de persistance, comme illustré à la **Fig. 6.9**, par une boucle de rétroaction positive affectant le dernier canal gagnant, nous pouvons filtrer efficacement des changements d'ordres moteurs intempestifs. A ce stade, il peut sembler que c'est cette introduction d'un filtrage temporel qui peut être à l'origine de la capacité du module MFI à diminuer le nombre de mouvements de tête en comparaison à un robot naïf. Cependant, le mécanisme de persistance ne sera utilisé que rarement, comme dans le cas limite où deux sources apparaissent au même moment et où l'évolution de l'apprentissage du M-SOM aboutit à une dynamique du critère q identique. Ce cas est rare et c'est bien l'analyse entière effectuée par le module MFI qui permet une modulation performante des mouvements de tête. Mais afin de s'assurer que cette persistance n'est qu'une petite partie de la modulation des mouvements de tête au sein du module MFI (le critère q étant le contributeur principal), nous testerons l'impact de cette persistance sur le nombre de mouvements de tête à la **Sec. 6.5.3**.

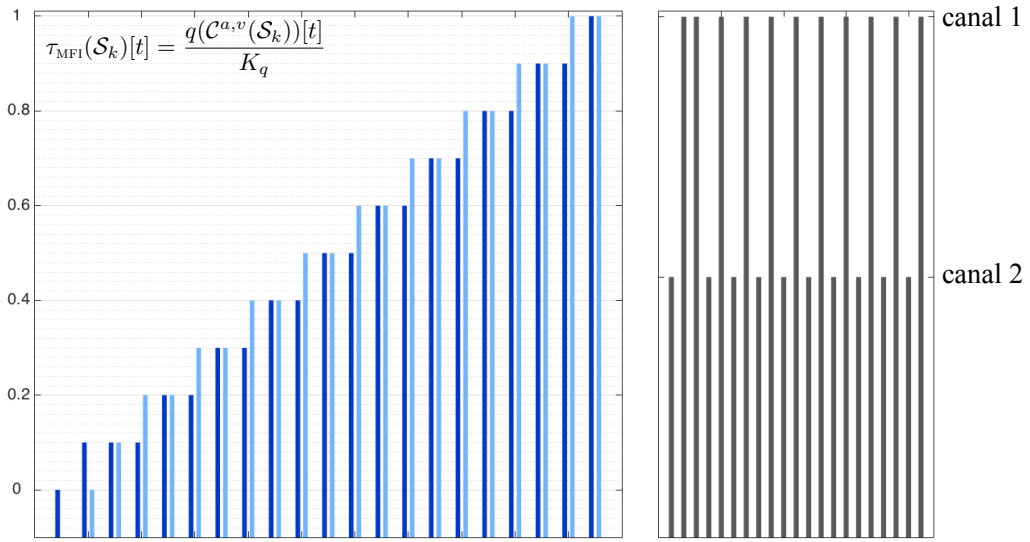


FIGURE 6.8 – ORDRES MOTEURS SANS PERSISTANCE — Scénario dans lequel deux sources sonores apparaissent quasiment au même moment et pour lesquelles le module MFI ne juge pas sa connaissance sur les catégories auxquelles elles appartiennent assez fiable, requérant ainsi un mouvement de tête vers chacune des deux. (*gauche*) activité du (*bleu foncé*) canal 1 et (*bleu clair*) canal 2 correspondant aux canaux d'activité des sources 1 et 2 respectivement. (*droite*) ordre moteur généré en conséquence du calcul du canal gagnant par l'**Eq. 6.10**.

La formalisation de la génération de l'ordre moteur grâce à la boucle ganglions de la base — thalamus — cortex permet de lier directement la confiance que le système a en sa connaissance de l'environnement avec l'attraction vers un ordre moteur particulier. De façon similaire à l'ordre moteur généré par le module DW, nous considérons ce mouvement moteur comme une motivation par la réduction de l'incertitude sur l'environnement.

Le module MFI a maintenant été présenté et sa formalisation a été décrite. La section suivante est ainsi dédiée aux différentes évaluations de ses multiples caractéristiques.

6.5 Résultats

AFIN de tester les performances du module MFI, l'environnement de simulation décrit à la **Sec. 4.3** a été utilisé. En plus du module MFI, deux autres systèmes ont été implémentés dans un but de comparaison. Ces deux systèmes additionnels effectuent une fusion des classifieurs directement à la sortie des experts d'identification, l'un incluant des mouvements de tête vers chaque nouvel source audiovisuelle apparaissant dans l'environnement, l'autre excluant les mouvements de tête.

D'autre part, contrairement au conditions expérimentales utilisées pour la validation du module DW, les sources audiovisuelles seront désormais placées tout autour du robot. Ainsi, nous allons pouvoir évaluer la capacité du module MFI à inférer une modalité lorsque celle-ci est absente. A noter que nous avons créé des scénarios

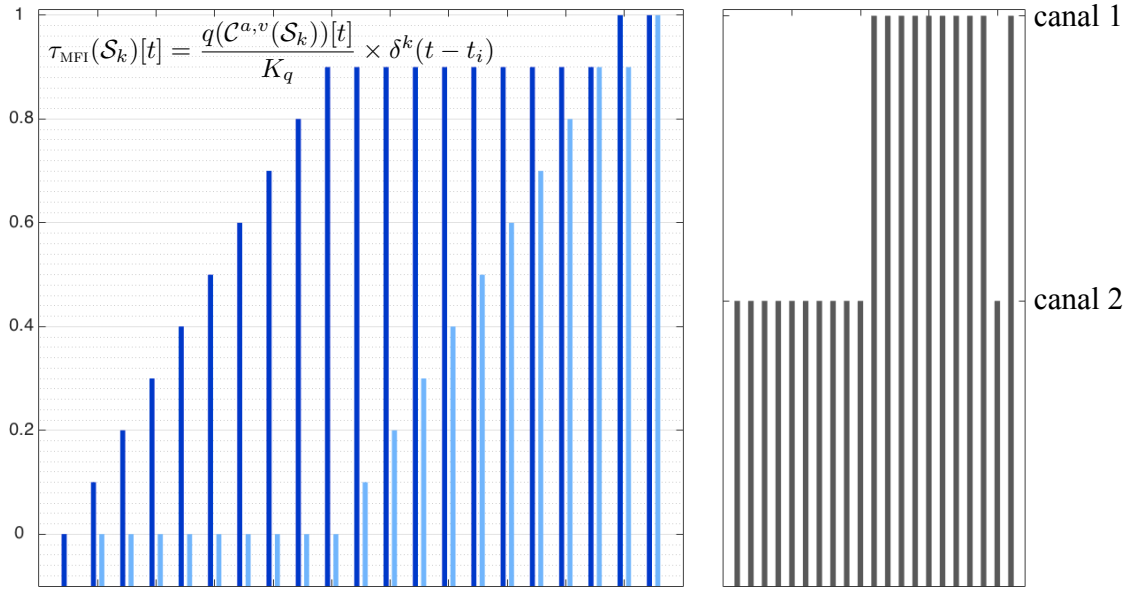


FIGURE 6.9 – ORDRES MOTEURS AVEC PERSISTANCE — Scénario identique à celui décrit à la **Fig. 6.8** mais cette fois-ci, l'ordre moteur généré se base sur un calcul du canal gagnant par l'**Eq. 6.13** incluant une persistance introduite par une boucle de rétroaction positive.

distincts pour chaque hypothèse à valider. Ces scénarios sont parfois très similaires, parfois même identiques. Cependant, nous avons tenu à recréer ces scénarios pour chaque condition de test afin de mettre en avant, indirectement, la robustesse du module.

La Sec. 6.5.1 consiste en l'évaluation des performances du module MFI sur la fusion de données et les capacités d'inférence de données lorsqu'une modalité est manquante.

La Sec. 6.5.2 présentera l'évaluation de l'impact du seuil K_q sur le comportement global du module MFI.

6.5.1 Classification audiovisuelle

Cette section présente les simulations effectuées pour tester l'hypothèse suivante :

Hypothèse 3. *La fusion de classifieurs opérée par le module MFI, via le M-SOM, est plus performante qu'une fusion en sortie directe des classifieurs.*

Cette hypothèse a été testée dans deux cas : le cas unisource et le cas multisource. Le cas unisource permet de mettre en valeur la capacité du module MFI à effectuer une fusion des données en fonction d'un taux d'erreur de sortie des experts $\varepsilon_{\mathcal{P}}$. A l'opposé, le cas multisource permettra de mettre en valeur la capacité du module

MFI a inférer les données manquantes issues des sources audiovisuelles hors de la portée du robot.

Nous rappelons ici les critères utilisés pour la validation de cette hypothèse :

- $\bar{\Gamma}^{a'}_{\text{MFI}}[t = T]$: taux de bonne classification en fin d'exploration calculé, pour chaque objet, à partir du moment où la source se met à émettre ;
- $\bar{\Gamma}^{a''}_{\text{MFI}}[t = T]$: taux de bonne classification en fin d'exploration calculé, pour chaque objet, à partir du moment où le module MFI effectue sa première tentative de fusion (s'il a accès à toutes les données) ou d'inférence (si une modalité est manquante) ;
- $\bar{\Gamma}_{\mathfrak{R}_n}[t = T]$: taux de bonne classification en fin d'exploration calculé, pour chaque objet, à partir d'un système de fusion directement effectuée en sortie des classifieurs, c'est-à-dire par sélection des composantes maximum du vecteur $\mathbf{P}^a[t]$ et $\mathbf{P}^v[t]$. A noter que le robot est ici omniscient : il a accès, tout le temps, à toutes les données audiovisuelles des sources présentes dans l'environnement (équivalent d'une inférence parfaite) ;
- $\bar{\Gamma}'_{\mathfrak{R}_n}[t = T]$: idem que $\bar{\Gamma}_{\mathfrak{R}_n}[t = T]$ mais cette fois-ci le robot $\mathfrak{R}_n$ n'est pas capable d'effectuer d'inférence. Ainsi le taux de bonne classification issu de la fusion directe des classifieurs n'est calculé que lorsque le robot a accès à toutes les données, c'est-à-dire lorsqu'il fait face à une source audiovisuelle.

6.5.1.1 Condition unisource

20 conditions de test ont été créées en faisant à chaque fois varier le taux d'erreur de classification par les experts d'identification et le nombre de sources présentes dans l'environnement. Ces conditions sont présentées au **Tab. 6.1**. Chaque condition a été testée cinq fois pour un total de 100 simulations.

Conditions de test 6.5.1.1 [7]						
No	$n_{\mathcal{S}}$	n_{sim}^{max}	T	Catégories présentes [8]	K_q	ε_p
1 à 5	3	1	500	1, 9	0.8	0.1, 0.3, 0.5, 0.7, 0.9
6 à 10	5	1	500	1, 9, 18	0.8	0.1, 0.3, 0.5, 0.7, 0.9
11 à 15	7	1	500	1, 9, 18, 28	0.8	0.1, 0.3, 0.5, 0.7, 0.9
16 à 20	10	1	500	1, 3, 9, 11, 18, 28, 41	0.8	0.1, 0.3, 0.5, 0.7, 0.9

TABLE 6.1 – Caractéristiques des 20 scénarios générés pour étudier la qualité de l'apprentissage effectuée par le module MFI en fonction de la complexité des scénarios. Chaque condition de test est répétée 5 fois, pour un total de 100 simulations.

Les résultats listés au **Tab. 6.2** montrent que le module MFI effectue une fusion quasiment systématiquement meilleure que celle faite directement à la sortie des experts d'identification : le ratio entre les deux types de fusion varie de 0.957 (légèrement en défaveur du module MFI) à 5.145 (largement en faveur du module MFI). Même pour des taux d'erreur très élevés, jusqu'à $\varepsilon_{\mathcal{P}} = 0.9$, le module MFI, grâce à la double fusion des données (intramodale et intermodale) qu'il effectue, parvient à

7. Se référer à la **Sec. 4.3** pour l'explication de ces notations.
8. Se référer au **Tab. 4.1** pour la liste des catégories audiovisuelles.

Résultats pour les conditions 6.5.1.1					
$\varepsilon_{\mathcal{P}}$	$n_{\mathcal{S}}$	$\bar{\Gamma}^{a'}_{\text{MFI}}[t=T]$	$\bar{\Gamma}^{a''}_{\text{MFI}}[t=T]$	$\bar{\Gamma}_{\mathfrak{R}_n}[t=T]$	ratio
0.1	3	0.927 (0.156)	0.931 (0.167)	0.899 (0.052)	1.034
	5	0.834 (0.235)	0.838 (0.242)	0.874 (0.034)	0.957
	7	0.847 (0.184)	0.851 (0.187)	0.850 (0.050)	0.999
	10	0.862 (0.186)	0.867 (0.195)	0.867 (0.031)	0.997
	moyenne	**0.867**	**0.871**	**0.872**	**0.996**
0.3	3	0.993 (0.020)	0.997 (0.022)	0.690 (0.029)	1.442
	5	0.834 (0.219)	0.839 (0.228)	0.678 (0.033)	1.234
	7	0.825 (0.251)	0.829 (0.259)	0.679 (0.062)	1.216
	10	0.888 (0.131)	0.893 (0.130)	0.685 (0.059)	1.299
	moyenne	**0.885**	**0.889**	**0.683**	**1.298**
0.5	3	0.923 (0.145)	0.929 (0.152)	0.471 (0.041)	1.965
	5	0.991 (0.032)	0.997 (0.028)	0.477 (0.048)	2.082
	7	0.843 (0.212)	0.847 (0.219)	0.497 (0.051)	1.699
	10	0.935 (0.098)	0.941 (0.102)	0.483 (0.023)	1.942
	moyenne	**0.923**	**0.928**	**0.482**	**1.920**
0.7	3	0.976 (0.077)	0.982 (0.018)	0.299 (0.039)	3.226
	5	0.975 (0.059)	0.980 (0.071)	0.296 (0.033)	3.298
	7	0.898 (0.104)	0.903 (0.113)	0.301 (0.054)	2.989
	10	0.965 (0.060)	0.972 (0.069)	0.291 (0.045)	3.321
	moyenne	**0.953**	**0.959**	**0.296**	**3.229**
0.9	3	0.587 (0.289)	0.590 (0.297)	0.114 (0.033)	5.145
	5	0.447 (0.177)	0.450 (0.184)	0.107 (0.015)	4.186
	7	0.450 (0.202)	0.453 (0.212)	0.098 (0.039)	4.571
	10	0.386 (0.219)	0.389 (0.238)	0.097 (0.043)	3.998
	moyenne	**0.467**	**0.470**	**0.104**	**4.504**

TABLE 6.2 – Taux de bonne classification pour l'ensemble des conditions de simulations présentées au **Tab. 6.1**. Chaque résultat est une moyenne sur les 5 répétitions de chaque conditions (avec l'écart-type entre parenthèse), pour un total de 100 simulations. Les valeurs sont arrondies à la troisième décimale.

obtenir des taux de bonne classification bien supérieurs à ceux du robot naïf : pour $\varepsilon_{\mathcal{P}} = 0.9$, la moyenne des $\bar{\Gamma}^{a}_{\text{MFI}}[t=T]$ sur les quatre environnements est de $46,7\%$ alors qu'elle n'est que de $10,4\%$ pour le robot naïf.

Enfin, ces résultats nous permettent également de voir que la robustesse de l'analyse du module MFI, testée *via* la répétition de chaque condition de test, n'est pas forcément dépendante de la qualité des données qu'il a à traiter. Pour $\varepsilon_{\mathcal{P}} = 0.1$, les écarts-type correspondant se situent dans le même ordre de grandeur que pour $\varepsilon_{\mathcal{P}} = 0.9$, tandis qu'ils sont globalement plus faibles pour $\varepsilon_{\mathcal{P}} = 0.5$ et $\varepsilon_{\mathcal{P}} = 0.7$. De façon similaire, nous n'observons pas de dépendance entre la robustesse du module MFI et le nombre de sources audiovisuelles présentes dans l'environnement : pour $n_{\mathcal{S}} = 5$, les écarts-type correspondant prennent des valeurs parfois plus faibles que

pour $n_{\mathcal{S}} = 3$, parfois plus fortes que pour $n_{\mathcal{S}} = 9$.

Tous ces résultats nous permettent de valider la qualité de la fusion effectuée dans le cas unisource, comparée à l'analyse de la sortie directe des experts d'identification. Nous allons maintenant effectuer les mêmes analyses mais dans le cas multisource.

6.5.1.2 Condition multisource

Nous avons ensuite évalué les performances du module MFI en condition multisource. Les environnements de test créés sont similaires à ceux de la condition unisource, la seule différence étant pour le paramètre n_{sim}^{max} que nous avons fait varier conjointement avec le nombre de sources présentes $n_{\mathcal{S}}$ de telle sorte que $n_{sim}^{max} = n_{\mathcal{S}}$. Les conditions de tests sont résumées au **Tab. 6.3**.

Conditions de test 6.5.1.2 [9]						
No	$n_{\mathcal{S}}$	n_{sim}^{max}	T	Catégories présentes [10]	K_q	ε_p
1 à 5	3	3	1000	1, 9	0.8	0.1, 0.3, 0.5, 0.7, 0.9
6 à 10	5	5	1000	1, 9, 18	0.8	0.1, 0.3, 0.5, 0.7, 0.9
11 à 15	7	7	1000	1, 9, 18, 28	0.8	0.1, 0.3, 0.5, 0.7, 0.9
16 à 20	10	10	1000	1, 9, 11, 18, 28, 41	0.8	0.1, 0.3, 0.5, 0.7, 0.9

TABLE 6.3 – Caractéristiques des 20 scénarios générés pour étudier la qualité de l'apprentissage effectuée par le module MFI en fonction de la complexité des scénarios. Chaque condition de test est répétée 5 fois pour un total de 100 simulations.

Tous les résultats de taux de bonne classification obtenus sur les 100 simulations sont rassemblés au **Tab. 6.4**. Ici encore, de façon globale, les taux obtenus par le module MFI sont tous, à l'exception d'une condition ($n_{\mathcal{S}} = 10$ et $\varepsilon_{\mathcal{P}} = 0.1$), au-dessus de ceux obtenus (i) par le robot omniscient et (ii) le robot naïf. Nous observons cependant une chute des taux du module MFI pour un taux d'erreur de 0.9% : pour la condition la plus extrême ($n_{\mathcal{S}} = 10$ et $\varepsilon_{\mathcal{P}} = 0.9$) le taux $\bar{\Gamma}^{a'}_{\text{MFI}}[t = T]$ tombe à 12.9% de classification correcte. Le robot naïf omniscient ne fait guère mieux, avec $\bar{\Gamma}_{\mathfrak{R}_n}[t = T] = 10.0\%$ de bonne classification et le robot naïf non omniscient, incapable d'effectuer une inférence de données, plonge à un taux de 1.9%. Malgré tout, le ratio entre le module et le robot naïf est encore en faveur du premier.

D'autre part, nous constatons que le module MFI est plus robuste dans le cas multisource que dans le cas unisource : 90% des écarts-type se situe en-dessous de 10% contre seulement 27.5% dans le cas unisource. Cette différence provient du rapport entre la valeur de la persistance temporelle de l'ordre moteur (cf. **Sec. 6.4**) et la durée durant laquelle les objets sont présents dans l'environnement (nous rappelons que les objets émettent durant une période alland de 15 à 35 trames, comme décrit à la **Sec. 4.3** et que la valeur de la persistance t_p a été fixée à 10 trames). En cas multisource, plusieurs sources sonores seront candidates, au temps t, à un mouvement de tête. Ainsi, une fois la période de persistance temporelle dépassée, le module MFI va pouvoir éventuellement déclencher un mouvement de tête vers

9. Se référer à la **Sec. 4.3** pour l'explication de ces notations.
10. Se référer au **Tab. 4.1** pour la liste des catégories audiovisuelles.

Résultats pour les conditions 6.5.1.2						
$\varepsilon_{\mathcal{P}}$	$n_{\mathcal{S}} \mid n_{sim}^{max}$	$\bar{\Gamma}^{a'}_{\text{MFI}}[t=T]$	$\bar{\Gamma}^{a''}_{\text{MFI}}[t=T]$	$\bar{\Gamma}_{\Re_n}[t=T]$	$\bar{\Gamma}'_{\Re_n}[t=T]$	ratio 1
0.1	3 | 3	0.960 (0.058)	0.982 (0.027)	0.894 (0.021)	0.503 (0.073)	1.086
	5 | 5	0.948 (0.036)	0.988 (0.025)	0.899 (0.012)	0.339 (0.039)	1.076
	7 | 7	0.909 (0.044)	0.960 (0.023)	0.893 (0.016)	0.264 (0.021)	1.046
	10 | 10	0.814 (0.045)	0.866 (0.047)	0.887 (0.018)	0.182 (0.014)	0.946
	moyenne	**0.907**	**0.949**	**0.893**	**0.322**	**1.038**
0.3	3 | 3	0.980 (0.046)	0.992 (0.020)	0.703 (0.042)	0.414 (0.055)	1.402
	5 | 5	0.951 (0.041)	0.987 (0.022)	0.692 (0.017)	0.265 (0.014)	1.399
	7 | 7	0.893 (0.051)	0.942 (0.028)	0.691 (0.014)	0.198 (0.017)	1.327
	10 | 10	0.805 (0.049)	0.883 (0.041)	0.689 (0.011)	0.145 (0.014)	1.224
	moyenne	**0.907**	**0.951**	**0.693**	**0.255**	**1.338**
0.5	3 | 3	0.956 (0.044)	0.973 (0.026)	0.493 (0.020)	0.280 (0.031)	1.955
	5 | 5	0.922 (0.078)	0.965 (0.043)	0.496 (0.021)	0.189 (0.034)	1.900
	7 | 7	0.853 (0.057)	0.899 (0.048)	0.492 (0.018)	0.145 (0.019)	1.779
	10 | 10	0.756 (0.054)	0.836 (0.042)	0.492 (0.018)	0.103 (0.010)	1.616
	moyenne	**0.871**	**0.918**	**0.493**	**0.179**	**1.812**
0.7	3 | 3	0.765 (0.075)	0.774 (0.087)	0.282 (0.030)	0.165 (0.028)	2.463
	5 | 5	0.713 (0.080)	0.737 (0.105)	0.294 (0.014)	0.120 (0.023)	3.298
	7 | 7	0.650 (0.101)	0.683 (0.133)	0.296 (0.016)	0.081 (0.012)	2.250
	10 | 10	0.505 (0.083)	0.550 (0.117)	0.293 (0.016)	0.064 (0.011)	1.801
	moyenne	**0.658**	**0.686**	**0.291**	**0.107**	**2.453**
0.9	3 | 3	0.211 (0.057)	0.213 (0.060)	0.092 (0.019)	0.054 (0.019)	2.303
	5 | 5	0.147 (0.054)	0.152 (0.064)	0.102 (0.012)	0.039 (0.007)	1.466
	7 | 7	0.163 (0.061)	0.174 (0.075)	0.100 (0.009)	0.031 (0.005)	1.679
	10 | 10	0.129 (0.042)	0.140 (0.066)	0.100 (0.009)	0.019 (0.006)	1.343
	moyenne	**0.162**	**0.169**	**0.098**	**0.035**	**1.697**

TABLE 6.4 – Taux de bonne classification pour l'ensemble des conditions de simulations présentées au **Tab. 6.1**. Chaque résultat est une moyenne sur les 5 répétitions de chaque conditions (avec l'écart-type entre parenthèse), pour un total de 100 simulations. Les valeurs sont arrondies à la troisième décimale.

une autre source. Dans le cas unisource en revanche, une fois la période de persistance dépassée, aucune autre source n'émet est aucun autre ordre moteur ne peut être généré. La conséquence de ce comportement est que l'apprentissage du lien audiovisuel sera légèrement plus difficile dans le cas unisource puisque le système sera confronté plus longtemps à des données correspondant à la même catégorie audiovisuelle entraînant ainsi des sous-réseaux sur-apprenant légèrement les catégories correspondantes. Dans le cas multisource en revanche, le fait d'avoir accès plus souvent à des données éventuellement issues d'objets appartenant à des catégories audiovisuelles différentes entraîne un apprentissage plus proche du paradigme SOM traditionnel dans lequel, à chaque itération, un vecteur de données est pris au hasard dans la matrice de données totale. Nous rappelons que cette méthode

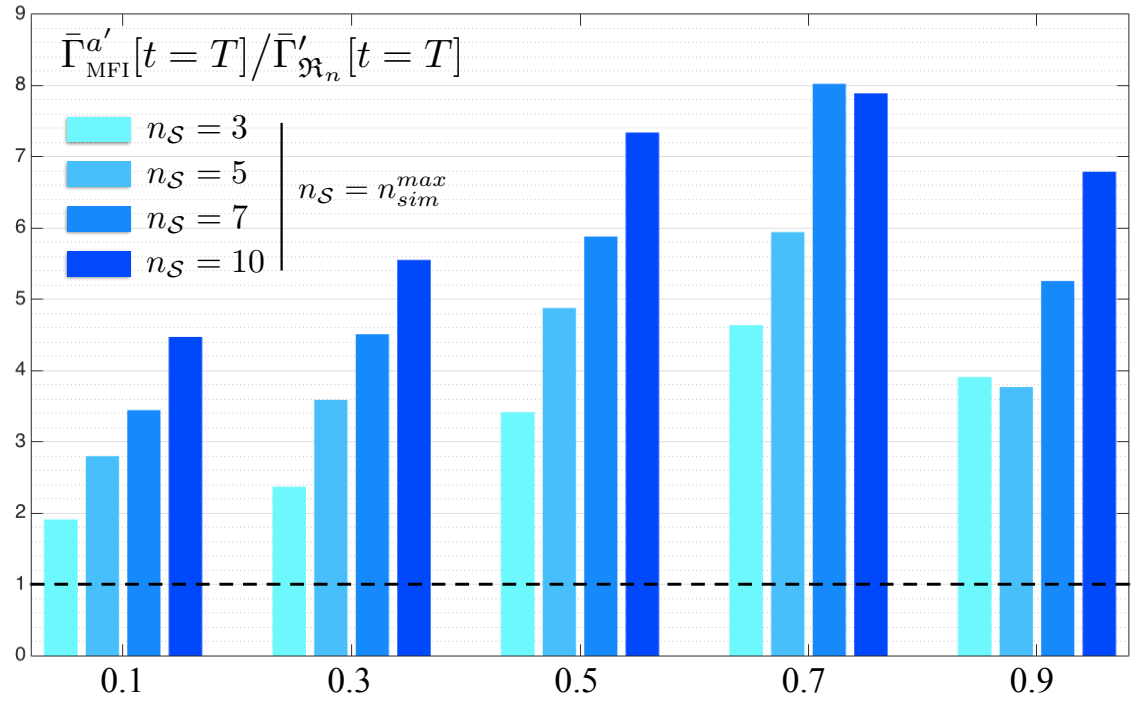


FIGURE 6.10 – RATIO ENTRE MFI ET ROBOT NAÏF — Taux de bonne classification audiovisuelle en fonction du nombre de sources audiovisuelles présentes dans l'environnement et du taux d'erreur des experts d'identification (cf. **Tab. 6.1**). Les histogrammes sont rassemblés par taux d'erreur $\varepsilon_{\mathcal{P}}$ simulé et chaque couleur représente le nombre de sources $n_{\mathcal{S}}$ aussi bien que le nombre maximum de sources n^{max}_{sim} émettant simultanément.

permet justement d'éviter le sur-apprentissage de certaines données et de garantir un apprentissage graduel et non biaisé par l'organisation temporelle des données à apprendre. Dans notre cas, cette différence de comportement entre cas unisource et cas multisource entraîne une légère baisse de la robustesse du module MFI exprimée par des écarts-type nettement inférieurs dans le second cas que dans le premier.

La **Fig. 6.10** présente le ratio entre le taux de bonne classification final $\bar{\Gamma}^{a'}_{\text{MFI}}[t=T]$ (où $a=\mathbf{a}'=1/[1,...,(t-t^1_{\mathcal{S}_j})+1]$, cf. **Eq. 4.10**) et le taux $\bar{\Gamma}'_{\mathfrak{R}_n}[t=T]$ du robot non doté de capacités d'inférence et ne pouvant ainsi qu'effectuer une fusion des données auxquelles il a accès. Cette figure nous permet de mettre en avant l'importance et l'intérêt de la capacité du module MFI à être capable d'inférer une modalité manquante : les ratios des taux de classification sont systématiquement en faveur du module MFI, allant de 1.908 pour la condition n°1, la plus simple, à 8.024 pour la condition n°15.

6.5.1.3 Discussion

Tous ces résultats nous permettent de valider l'**Hyp. 3** selon laquelle la fusion opérée par le module MFI est meilleure que celle effectuée en sortie directe des classifieurs. En effet, quelle que soit la condition de test (à deux exceptions près) et dans le cas unisource ou celui multisource, le module MFI surpasse largement les résultats obtenus par le robot naïf. De plus, le module parvient à gérer efficacement

les erreurs de classification générées par les classifieurs, même dans des cas très extrêmes. Ces performances sont majeures car les catégories audiovisuelles émises par le module MFI vont être au cœur du calcul de la Congruence effectué par le module DW. Ainsi, la qualité des données reçues est indispensable à l'élaboration d'une réaction attentionnelle pertinente.

Ces bons résultats sont possibles grâce à l'utilisation des mouvements de tête pour accéder à une information nécessaire au module MFI pour l'apprentissage du M-SOM. Mais ceux-ci sont en conflit avec ceux générés par le module DW. Il est ainsi important d'étudier le comportement du module MFI du point de vue de la génération de ces mouvements. La section suivante consiste ainsi en l'évaluation de la capacité du module MFI à inhiber des mouvements de tête, comportement qui, en conjonction avec les taux de bonne classification, est la preuve de la convergence de l'apprentissage effectué par le module.

6.5.2 Evaluation du critère K_q

Le critère K_q permet de moduler la qualité de l'apprentissage en agissant sur la rapidité à laquelle le module MFI fait confiance en sa capacité à inférer une modalités, par comparaison avec le critère $q(\widehat{\mathcal{C}}^{a,v})$. Ce critère a également un effet sur les mouvements de tête générés puisque ceux-ci sont utilisés pour accéder aux données visuelles lorsqu'elles sont manquantes et que le module MFI cherche à apprendre. Les deux sections suivantes présentent les résultats sur la l'évaluation de ces deux effets.

6.5.2.1 Influence sur l'apprentissage

Hypothèse 4. *Le critère K_q permet de moduler le comportement du module MFI et sa valeur a un impact sur la qualité de l'apprentissage.*

Nous avons créé deux environnements de test extrêmes, un unisource et un multisource, présentés au **Tab. 6.5**, et pour lequel nous avons fait varier la valeur de K_q entre 0.1 et 1.0, sa valeur maximale (une valeur de $K_q = 0.0$ inhibera tous les mouvements de tête et n'a donc pas été testée).

Conditions de test 6.5.2.1 [11]						
No	$n_{\mathcal{S}}$	n_{sim}^{max}	T	Catégories présentes [12]	K_q	ε_p
1 à 4	10	1	1000	1, 9, 18, 29, 32, 36	0.1, 0.4, 0.7, 1.0	0.3
5 à 8	10	5	1000	1, 9, 18, 29, 32, 36	0.1, 0.4, 0.7, 1.0	0.3

TABLE 6.5 – Caractéristiques des 8 scénarios générés pour étudier l'impact du critère K_q sur le comportement du module MFI. Chaque condition a été testée 5 fois pour un total de 40 simulations.

11. Se référer à la **Sec. 4.3** pour l'explication de ces notations.
12. Se référer au **Tab. 4.1** pour la liste des catégories audiovisuelles.

Résultats pour les conditions 6.5.2.1						
K_q	$n_{\mathcal{S}} \mid n_{sim}^{max}$	$\bar{\Gamma}^{a'}_{\text{MFI}}[t=T]$	$\bar{\Gamma}^{a''}_{\text{MFI}}[t=T]$	$\bar{\Gamma}_{\mathfrak{R}_n}[t=T]$	$\bar{\Gamma}'_{\mathfrak{R}_n}[t=T]$	ratio
	10 | 1	0.944 (0.054)	0.946 (0.054)	0.665 (0.024)		1.421
	10 | 10	0.578 (0.392)	0.606 (0.419)	0.699 (0.010)	0.145 (0.001)	0.846
0.1	**moyenne**	**0.761**	**0.776**	**0.682**		
	10 | 1	0.959 (0.027)	0.961 (0.027)	0.672 (0.502)		1.428
	10 | 10	0.859 (0.025)	0.879 (0.039)	0.691 (0.008)	0.143 (0.007)	1.256
0.4	**moyenne**	**0.909**	**0.920**	**0.681**		
	10 | 1	0.979 (0.011)	0.982 (0.011)	0.672 (0.042)		1.459
	10 | 10	0.824 (0.030)	0.908 (0.051)	0.696 (0.007)	0.141 (0.002)	1.244
0.7	**moyenne**	**0.901**	**0.945**	**0.684**		
	10 | 1	0.991 (0.004)	0.995 (0.004)	0.674 (0.033)		1.473
	10 | 10	0.792 (0.069)	0.879 (0.082)	0.689 (0.008)	0.139 (0.006)	1.213
1.0	**moyenne**	**0.891**	**0.937**	**0.681**		

TABLE 6.6 – Taux de bonne classification pour l'ensemble des conditions de simulations présentées au **Tab. 6.5**. Chaque résultat est une moyenne sur les 5 répétitions de chaque conditions (avec l'écart-type entre parenthèse), pour un total de 20 simulations. Les valeurs sont arrondies à la troisième décimale.

Le **Tab. 6.6** présente les mesures de la qualité de catégorisation audiovisuelle du module MFI et du robot naïf (nous rappelons que K_q n'a aucun effet sur le robot naïf). Concernant le cas unisource, nous observons tout d'abord de très bonnes performances globales, systématiquement supérieures au robot naïf omniscient et ce, même dans le cas où $K_q = 0.1$, sa valeur minimum, c'est-à-dire une valeur pour laquelle après neuf mauvaises inférences d'une catégorie audiovisuelle, si la dixième est correcte, le module MFI considérera qu'il a suffisamment appris cette catégorie pour inhiber un mouvement de tête. Deuxièmement, nous observons une augmentation des taux de bonne classification du module MFI conjointement avec l'augmentation de K_q. Cependant, le cas unisource n'est pas vraiment pertinent ici : quelle que soit la valeur de K_q, une fois le mouvement de tête généré, le module MFI apprend les données audiovisuelles sans que la catégorie d'aucune autre source audiovisuelle ne soit à inférer. Le temps durant lequel le robot fait face à une source donnée est suffisant pour apprendre correctement la catégorie concernée et même si la première inférence s'était trouvée fausse (ce qui est hautement probable), la seconde, plus tard, aurait de grandes chances de se trouver juste. Ainsi, K_q n'a quasiment aucun effet sur les performances du module MFI en condition unisource.

Le cas multisource est quant à lui plus intéressant : nous voyons d'importantes disparités entre les taux de bonne classification en fonction de K_q. Notamment, nous observons une chute de ce taux pour $K_q = 0.1$ et un ratio ainsi en défaveur du module MFI. Ceci s'explique pour la raison inverse : à chaque pas de temps, jusqu'à dix sources peuvent émettre simultanément, chacune de ces sources pouvant requérir, surtout en début de simulation, un mouvement de tête. Il est ainsi probable qu'alors que le robot faisait face à une source, une autre source ait requis un mouvement de tête, sélectionnée par le processus décrit à la **Sec. 6.4**, avant que le module MFI

ait été capable d'apprendre correctement la catégorie de la première source. Nous voyons en revanche que pour $K_q = 0.4$ et $K_q = 0.7$, les taux de bonne classification remontent et dépassent significativement les taux du robot naïf, omniscient ou non. Enfin, nous observons que pour $K_q = 1.0$, sa valeur maximale, c'est-à-dire pour laquelle aucun mouvement de tête ne peut être inhibé, nous constatons une légère baisse des taux de bonne classification. Cette baisse est causée par le fait que le système ne s'arrête jamais d'apprendre et s'explique par la façon dont nous avons modifié l'incrémentation de l'itération d'apprentissage. Comme décrit à la **Sec. 6.2.3.2**, nous avons inversé l'incrémentation de l'itération conditionnant la force de l'apprentissage : l'apprentissage des premières données n'est que très peu propagé dans les sous-réseaux tandis que l'apprentissage des dernières données est au contraire propagé au maximum. Ainsi, une fois que l'ensemble des catégories a été perçu durant un temps assez long (temps qui sera étudié à la **Sec. 6.5.4**), chaque nouvelle donnée audiovisuelle entraînant une étape d'apprentissage aura un effet non négligeable sur chacun des deux sous-réseaux. De plus, il faut coupler à ce phénomène, le fait que les données sont acquises de façon temporellement organisées : durant n trames temporelles, des données concernant la même catégorie seront apprises par le réseau. Le taux d'apprentissage ainsi que la fonction de voisinage étant à leur maximum, l'impact de la réorganisation sera d'autant plus fort, éventuellement changeant des zones définies représentant d'autres catégories. En conséquence, les sous-réseaux seront plus sujets à des modifications, augmentant par là le nombre d'erreurs de classification par le module MFI. Mais même dans ce cas extrême, nous observons des taux de bonne classification pour $K_q = 1.0$ supérieurs au robot naïf omniscient.

Nous pouvons donc valider l'**Hyp. 4** : le critère K_q module bien la qualité de l'apprentissage.

6.5.2.2 Influence de K_q sur les mouvements de tête

Hypothèse 5. *Le critère K_q a un effet sur le nombre de mouvements de tête générés par le MFI.*

La **Fig. 6.11** illustre l'impact du critère K_q sur le nombre de mouvements de tête générés. Ce nombre varie directement avec la valeur de K_q : plus celui-ci est élevé, plus il y a de mouvements de tête générés. Par ailleurs, nous observons que même lorsque $K_q = 1,0$, c'est-à-dire lorsqu'il est à sa valeur maximale, le nombre de mouvements de tête est plus faible que pour le robot naïf. Cependant, ce comportement n'est valable qu'en condition multisource. En effet, en cas unisource, $\mathfrak{R}_{\text{MFI}}$ aurait le même comportement que $\mathfrak{R}_n$ puisque ne faisant jamais confiance en ses capacités d'inférence, il générerait des commandes motrices vers chaque événement survenant dans l'environnement, exactement comme le robot naïf donc. Dans le cas multisource, la valeur plus faible du nombre de mouvements de tête observée même lorsque $K_q = 1.0$ s'explique par la façon dont les ordres moteurs sont générés et particulièrement le processus de persistance de l'ordre moteur implémenté comme une boucle de rétrocontrôle positive affectant le dernier canal gagnant.

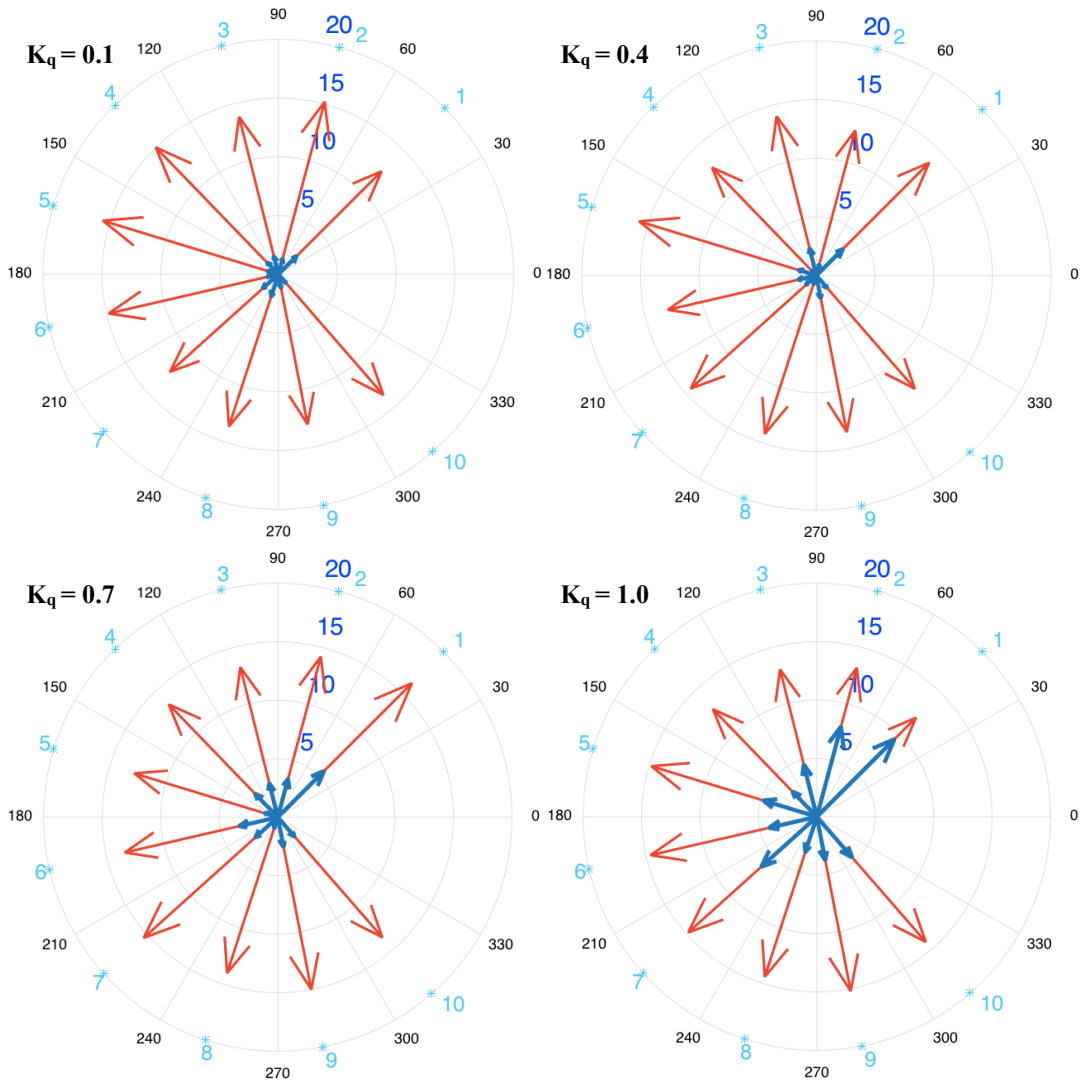


FIGURE 6.11 – INFLUENCE DE K_q SUR LES MOUVEMENTS DE TÊTE — Le critère K_q a un impact sur la vitesse à laquelle le M-SOM converge et donc, indirectement, sur le nombre de mouvements de tête générés par le module MFI. Chaque flèche représente le nombre de mouvements de tête générés vers la position des sources. (*flèches rouges*) mouvements de tête générés par le robot naïf, (*flèches bleues*) mouvements de tête générés par le module MFI. La longueur des flèches représentent le nombre de mouvements générés vers chacune des sources.

Nous pouvons donc valider l'**Hyp. 5** : le critère K_q a un effet sur le nombre de mouvements de tête.

6.5.2.3 Discussion

Le critère K_q est d'importance majeure dans le comportement du module MFI en cela qu'il permet de moduler la dynamique globale de l'apprentissage et, avec elle, le nombre de mouvements de tête générés. Que ce soit dans le cas où $K_q = 0,1$ ou dans celui où il est maximal, les taux de bonne classification $\bar{\mathbf{\Gamma}}^{a'}_{\text{MFI}}$ et $\bar{\mathbf{\Gamma}}^{a''}_{\text{MFI}}$ sont systématiquement supérieurs aux taux $\bar{\mathbf{\Gamma}}_{\mathfrak{R}_n}$ et $\bar{\mathbf{\Gamma}}'_{\mathfrak{R}_n}$. Cela implique que ce critère

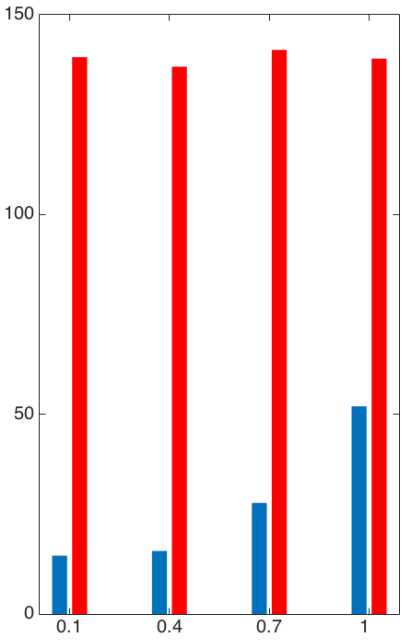


FIGURE 6.12 – INFLUENCE DE K_q SUR LES MOUVEMENTS DE TÊTE — Le critère K_q a un impact sur la vitesse à laquelle le M-SOM converge et donc, indirectement, sur le nombre de mouvements de tête générés par le module MFI. Cette figure est une représentation additionnelle des résultats présentés à la **Fig. 6.11**. Chaque paire d'histogrammes correspond à la moyenne des résultats obtenus après 5 simulations du scénario 1 et en faisant varier la valeur du critère $K_q = [0.1, 0.4, 0.7, 1.0]$. (*rouge*) nombre de mouvements de tête pour le robot naïf (nombre variable car les scénarios contiennent une part d'aléatoire lors de leur génération, cf. **Sec. 4.3**), (*bleu*) pour le module MFI.

n'est pas un seuil utilisé pour rendre le système particulièrement performant mais bien un critère permettant de décider quel comportement le robot doit adopter en fonction des situations dans lesquelles il se situe. Notamment, au sein de l'architecture TWO!EARS, le robot dispose d'un mode S&R qui peut se déclencher au cours de l'exploration d'un environnement en fonction des entités audiovisuelles perçues (cette décision est indépendante du modèle HTM et correspond à la réaction à un ensemble de règles de comportement données *a priori* au robot). Nous pouvons utiliser ce mode afin de modifier la valeur de K_q en temps réel et ainsi doter le robot d'un comportement évolutif et adaptatif.

D'autre part, nous voyons que même lorsque K_q est à sa valeur maximum, le nombre de mouvements de tête est largement inférieur au robot naïf. Nous rappelons que les décisions motrices sont modélisées grâce à un algorithme inspiré du modèle GPR permettant de gérer des décisions motrices contradictoires de façon performante, algorithme jouant un rôle dans cette diminution observée systématiquement. La section suivante est dédiée au test du phénomène de *persistance temporelle* faisant partie de l'algorithme de décision de l'ordre moteur gagnant.

6.5.3 Etude de la persistance

Hypothèse 6. *La persistance temporelle ne joue qu'un rôle mineur dans le nombre de mouvements de tête générés par le module MFI.*

Nous avons détaillé à la **Sec. 6.4** la façon dont les ordres moteurs sont générés à l'aide du calcul de l'activité des sources sonores candidates pour la génération d'une commande motrice. En particulier, nous avons introduit le processus de persistance temporelle implémentée comme une boucle de rétroaction positive permettant d'éviter la génération d'ordres contradictoires intempestive par favorisation du dernier

canal gagnant, exprimé par l'**Eq. 6.13**. Nous avions alors stipulé que l'introduction de cette persistance n'affectait pas de façon significative la modulation des mouvements de tête et n'était pas le facteur majoritaire dans la diminution du nombre de ces mouvements en comparaison au robot naïf. Cette section présente donc le test de l'**Hyp. 6** par un ensemble de simulations effectuées dans un environnement complexe et selon six conditions de test différentes (chacune étant effectuée cinq fois en tout). Chaque condition est caractérisée par une durée de la persistance temporelle : de $t_p = 1$ pas de temps à $t_p = 25$ pas de temps (conditions détaillées au **Tab. 6.7**. A titre de comparaison, cela correspondrait à des durées, sur le vrai robot, allant de 500 ms à $12,5$ s. Mais tester des durées aussi longues — et quelque peu absurdes dans des conditions réalistes — permet de véritablement prouver que ce phénomène n'est pas un facteur déterminant dans le nombre de mouvements de tête générés par le module MFI.

Conditions de test 6.5.3 [13]						
n°	$n_{\mathcal{S}}$	n_{sim}^{max}	T	Catégories présentes [14]	t_p	ε_p
1 à 6	8	5	500	1, 9, 15, 18, 21, 28, 41	1, 5, 10, 15, 20, 25	0.3

TABLE 6.7 – Caractéristiques des 6 conditions de tests générés dans un environnement complexe afin d'étudier l'impact de la valeur de la persistance t_p dans le nombre de mouvements de tête générés par le module MFI. Chaque condition a été répétée 5 fois.

Le **Fig. 6.13** illustre les résultats obtenus et nous permettent premièrement d'observer deux phénomènes :

1. il existe une dépendance entre le nombre de mouvements de tête générés et la persistance : il y a une diminution d'environ $13,6\%$ entre une persistance de valeur $t_p = 1$ et une persistance de valeur $t_p = 25$. Ainsi, plus la persistance est élevée, plus le nombre de mouvements de tête tend à diminuer.
2. une persistance plus élevée tend à diminuer la variabilité du nombre de mouvements de tête (barres d'erreurs de la **Fig. 6.13**) : d'un écart-type de $9,3$ mouvements à $t_p = 1$, nous arrivons à $4,2$ mouvements pour $t_p = 25$. Plus précisément, la persistance permet de rendre le module MFI légèrement plus constant (nous rappelons que chaque répétition d'une condition comporte une part d'aléatoire, notamment dans le décours temporel de chaque source sonore).

Mais ce que nous retiendrons de ces résultats est surtout le fait que même pour une valeur de persistance à $t_p = 1$, c'est-à-dire lorsqu'il n'y a aucune persistance, le module MFI est malgré tout en mesure de diminuer significativement le nombre de mouvements de tête en comparaison du robot naïf (et en son système attentionnel exclusivement basé sur une forme de motivation par la Nouveauté), ce qui valide l'**Hyp. 6**.

A ce stade, la question de l'intérêt de la persistance peut apparaître. La **Fig. 6.14** présente les mouvements de tête générés par le module MFI en fonction de la valeur de t_p et dans un scénario extrêmement simplifié (volontairement), au décours temporel déterminé préalablement à des fins illustratives. De plus, pour aboutir à cet

13. Se référer à la **Sec. 4.3** pour l'explication de ces notations.
14. Se référer au **Tab. 4.1** pour la liste des catégories audiovisuelles.

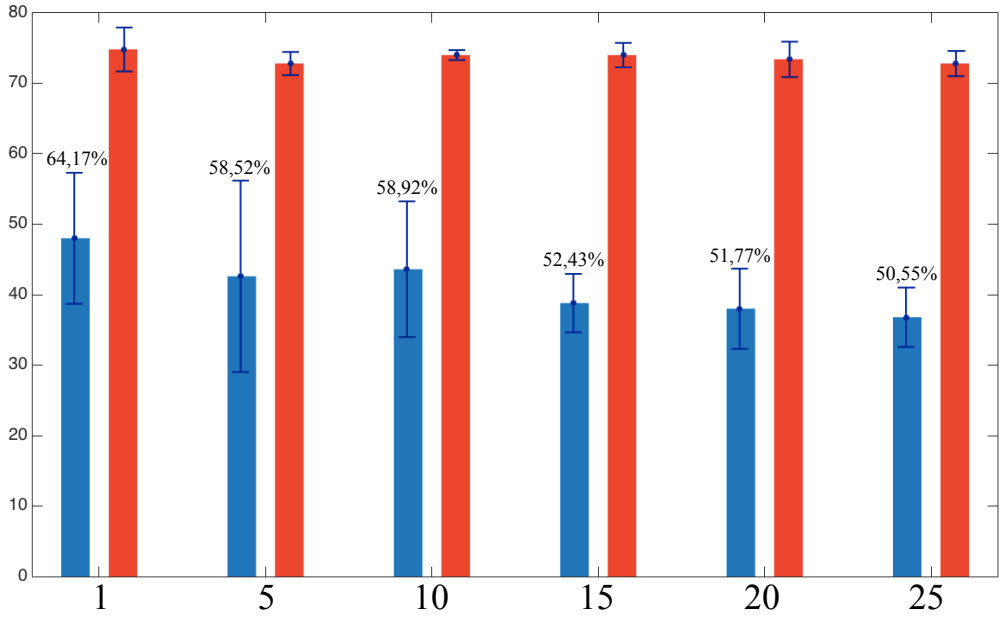


FIGURE 6.13 – VALEUR DE LA PERSISTANCE TEMPORELLE ET MOUVEMENTS DE TÊTE (I) — Résultats des simulations effectuées pour étudier l'impact de la durée de la persistance temporelle introduite à la **Sec. 6.4** sur le nombre de mouvements de tête générés par (*histogrammes bleus*) le module MFI, (*histogrammes rouges*) le robot naïf. Les données sont organisées en fonction de la valeur de la persistance t_p. Les pourcentages représentent la diminution du nombre de mouvements de tête grâce au module MFI par rapport au robot naïf.

---

extrême, nous avons enlevé au module MFI sa capacité d'inhiber des mouvements de tête. Nous observons ici l'intérêt de la persistance temporelle : lorsque $t_p = 1$, c'est-à-dire lorsque la persistance est inexistance, le système de génération d'ordres moteurs peut se retrouver « coincé » entre deux (ou plusieurs) décisions de commandes motrices. Rappelant la formalisation de la **Sec. 6.4**, les canaux représentant les deux sources *male speech* et *female singing* oscillent entre « gagnant / perdant / gagnant / perdant » etc. En revanche, dès lors que la boucle de rétroaction positive est introduite, ce phénomène d'oscillation disparaît et on observe un filtrage des ordres moteurs évitant le comportement observé pour $t_p = 1$. Nous observons également un effet pervers d'une trop grande persistance : pour $t_p = 25$, la première apparition de $\mathcal{S}_2$ (*female speech*) est complètement ignorée, au profit de $\mathcal{S}_1$ (*male speech*). Cette valeur de persistance est donc à prendre avec précaution puisqu'elle peut aboutir à un comportement contraire à celui que nous recherchons : apprendre le mieux possible l'environnement en cours d'exploration.

Il serait ici compréhensible de se poser la question suivante : pourquoi tenter d'éviter la situation observée lorsque $t_p = 1$? La réponse peut sembler évidente puisque si nous nous imaginons à la place du robot devant tourner sa tête incessamment d'un point à l'autre de l'espace, nous aurions sûrement du mal à tenir longtemps et à parvenir à garder une compréhension correcte de l'environnement. Mais considérant une plateforme robotique, générer un grand nombre de mouvements de tête n'est pas forcément problématique. Mettant de côté la pertinence comportementale d'un robot réagissant comme dans le cas où $t_p = 1$, il y a un intérêt majeur à éviter

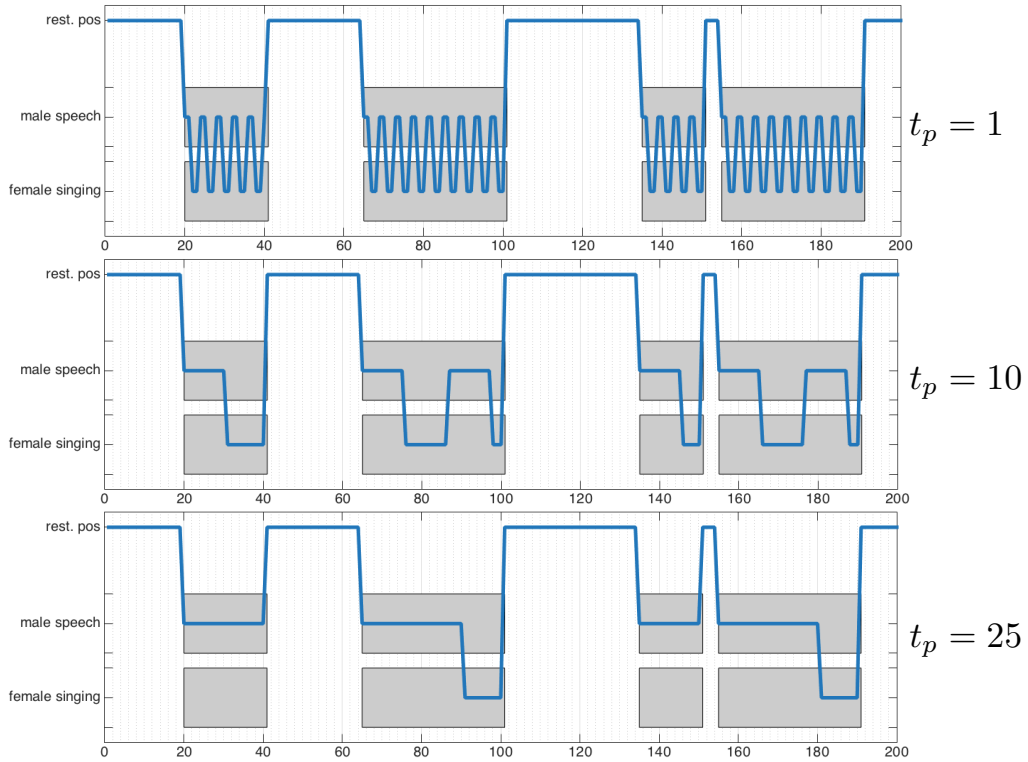


FIGURE 6.14 – VALEUR DE LA PERSISTANCE TEMPORELLE ET MOUVEMENTS DE TÊTE (II) — Résultats des simulations effectuées pour étudier l'impact de la durée de la persistance temporelle introduite à la **Sec. 6.4** sur le nombre de mouvements de tête générés par (*histogrammes bleus*) le module MFI, (*histogrammes rouges*) le robot naïf.

ce comportement : parvenir à stabiliser la représentation que le robot a de son environnement. Prenons par exemple le calcul des indices binauraux. Les experts de localisation et d'identification de TWO!EARS réalisent ces calculs sur des trames de 500 ms. Au sein de cette trame d'une demi-seconde, il peut y avoir beaucoup de variabilité dans la scène audiovisuelle, que ce soit par les mouvements du robot en train naviguer dans l'environnement ou par l'apparition de nouvelles sources sonores au milieu de cette trame et venant gêner la localisation de la source précédente. En ajoutant un degré de liberté au robot, par le cou permettant une rotation de la tête, nous ajoutons également un degré de complexité pouvant éventuellement gêner l'analyse de la scène audio. D'autre part, dans des systèmes intégrant des algorithmes de traques audio ou visuelle, une génération intempestive de mouvements de tête va également grandement complexifier ces tentatives de suivi d'objets. Ainsi, bien qu'un mouvement de tête occasionne une modification volontaire et maitrisée des données acoustiques (ou visuelles) perçues qui, comme nous l'avons vu précédemment, peuvent justement être utilisées pour améliorer significativement l'analyse de la scène audio, par exemple, il est nécessaire de les générer avec modération. La valeur de persistance de $t_p = 10$ que nous avons choisie permet d'effectuer un filtrage temorel moyen des commandes motrices conférant au robot une certaine stabilité.

6.5.4 Nombre d'itérations d'apprentissage

Le nombre d'itérations d'apprentissage d'un réseau de type SOM, comme le M-SOM, est un paramètre déterminant dans la qualité de l'apprentissage ainsi que dans le temps de convergence de l'algorithme. Contrairement à l'utilisation traditionnelle des cartes auto-adaptatives, le M-SOM ne dispose pas de l'ensemble des données : il les récupère trame par trame et doit parvenir à les catégoriser en temps réel. Deux possibilités dans ce contexte :

- garder un nombre d'itérations élevé afin de tirer un maximum de profit de l'acquisition d'une trame audiovisuelle ;
- réduire ce nombre d'itérations afin d'accélérer le processus d'apprentissage.

Mais au-delà de la rapidité d'exécution de la diminution du temps de convergence, le nombre d'itérations est aussi à déterminer en considérant la modification du paradigme d'apprentissage effectué pour le M-SOM. En effet, les données à traiter ici ont des caractéristiques particulières :

- elles sont obtenues séquentiellement ;
- deux vecteurs de données consécutifs correspondent la plupart du temps à un même objet.

Or, l'algorithme d'apprentissage des réseaux de type SOM impose le choix au hasard d'un vecteur au sein de la matrice entière de données à apprendre, et ce afin d'éviter que la partie du réseau codant cette catégorie ne se propage trop, jusqu'à envahir littéralement tout le réseau. Cette propagation résulterait en la grande difficulté — jusqu'à l'impossibilité — du réseau à intégrer de nouvelles données n'appartenant pas à la catégorie sur-apprise.

Une des pistes suivies au début de l'implémentation du module MFI a été de toujours garder en mémoire tous les vecteurs observés et de réinitialiser, à chaque nouvelle trame, le M-SOM entier afin de réapprendre à partir de zéro et suivant l'algorithme traditionnel d'apprentissage. Cependant, cette solution est (i) très « lourde », tant au niveau computationnel que conceptuel, et (ii) non-pertinente car impliquant que le système a une mémoire infinie. Ainsi, prenant en compte toutes ces données, nous avons choisi de n'apprendre que la dernière trame audiovisuelle perçue en essayant de limiter au maximum le nombre d'itérations d'apprentissage afin de ne pas faire converger le réseau trop rapidement vers des solutions locales, empêchant l'émergence de nouvelles catégories audiovisuelles.

Les expériences effectuées ici cherchent à démontrer l'hypothèse suivante :

Hypothèse 7. *Le nombre d'itérations d'apprentissage n_{it} peut être réduit à une valeur très proche de* 1 *tout en conservant :*

- *une haute performance dans l'apprentissage des catégories audiovisuelles,*
- *une convergence rapide,*
- *un temps computationnel bas.*

Afin de tester cette hypothèse, nous avons conduit une série de simulations en environnements relativement complexes et en faisant varier le nombre d'itérations d'apprentissage. De plus, la complexité des données à apprendre étant fonction de la

qualité des sorties des experts, chaque valeur de n_{it} a été testée en faisant également varier le taux d'erreur $\varepsilon_{\mathcal{P}}$. Les conditions sont présentées au **Tab. 6.8**.

Conditions de test 6.5.4 [15]						
n°	$n_{\mathcal{S}}$	n_{sim}^{max}	T	Catégories présentes [16]	n_{it}	ε_p
1 à 4	5	5	1000	1, 9, 18, 21	1	0.0, 0.25, 0.50, 0.75
5 à 8	5	5	1000	1, 9, 18, 21	10	0.0, 0.25, 0.50, 0.75
9 à 12	5	5	1000	1, 9, 18, 21	100	0.0, 0.25, 0.50, 0.75
13 à 16	5	5	1000	1, 9, 18, 21	1000	0.0, 0.25, 0.50, 0.75

TABLE 6.8 – Caractéristiques des 16 scénarios générés pour étudier l'impact du nombre d'itérations d'apprentissage du M-SOM sur la qualité de la classification audiovisuelle. Chaque scénario a été répété 5 fois en tout, pour un total de 80 simulations.

Le seuil K_q a été fixé à $0,8$ pour toutes les simulations afin de pousser le système a générer des mouvements de tête (pour rappel, plus le seuil K_q est élevé, moins le système fait confiance en son inférence). Enfin, étant donné que l'initialisation des vecteurs de poids du M-SOM est une étape importante de l'apprentissage (comme dans tout réseau de type SOM), une initialisation a été faite lors de la première simulation et a servi pour toutes les simulations suivantes.

Le **Tab. 6.9** présente les taux de bonne classification audiovisuelle moyen en fin de simulation, en fonction de n_{it} et $\varepsilon_{\mathcal{P}}$. Les taux sont très proches : le minimum est à 91.66 % pour ($n_{it} = 1, \varepsilon_{\mathcal{P}} = 0.25$) et le maximum à 95.70 % pour ($n_{it} = 10, \varepsilon_{\mathcal{P}} = 0.0$), soit une différence de 4.04 %. En moyennant les taux de bonne classification selon le taux d'erreur $\varepsilon_{\mathcal{P}}$, le minimum est à 92.63 % pour $n_{it} = 1$ et le maximum est à 95.67 % pour $n_{it} = 10$, soit 3.04 % de différence.

Résultats pour les conditions 6.5.4						
n_{it}	$\bar{\mathbf{\Gamma}}^{a'}_{\text{MFI}}[t = T]$ (%)				Moyenne	Ecart-type
	$\varepsilon_{\mathcal{P}} = 0.0$	$\varepsilon_{\mathcal{P}} = 0.25$	$\varepsilon_{\mathcal{P}} = 0.50$	$\varepsilon_{\mathcal{P}} = 0.75$		
1	93.81 %	91.66 %	92.57 %	92.48 %	92.63 %	0.0089
10	95.70 %	95.68 %	95.65 %	95.64 %	**95.67** %	< 0.001
100	94.69 %	94.70 %	94.06 %	94.71 %	94.54 %	0.0032
1000	94.67 %	94.64 %	94.61 %	94.09 %	94.50 %	0.0028

TABLE 6.9 – NOMBRE D'ITÉRATIONS DE L'APPRENTISSAGE ET QUALITÉ DE L'APPRENTISSAGE — Influence du nombre d'itérations sur la qualité de la classification audiovisuelle. Expériences réalisées en environnements simulés avec différents taux d'erreur $\varepsilon_{\mathcal{P}}$ des experts d'identification.

Deux observations intéressantes sont à faire ici :

- en multipliant par 1000 le nombre d'itérations d'apprentissage, le taux de bonne classification audiovisuelle n'augmente que d'environ 3% ;
- un plus grand nombre d'itérations ne signifie pas forcément, dans notre cas, un meilleur apprentissage.

15. Se référer à la **Sec. 4.3** pour l'explication de ces notations.
16. Se référer au **Tab. 4.1** pour la liste des catégories audiovisuelles.

Ce résultat est principalement dû à la façon dont nos données sont collectées et traitées par le module MFI et le M-SOM en particulier. Nous rappelons que le M-SOM n'apprend qu'un vecteur à la fois (celui issu de la dernière perception audiovisuelle du robot) et que le numéro de l'itération au temps t dépend du nombre de trames observées pour l'objet o_j considéré (cf. **Sec. 6.2.3.2**). De plus, le module MFI permet de se focaliser sur une source audivisuelle grâce à la façon dont sont générés les ordres moteurs (cf. **Sec. 6.4**), entraînant ainsi la perception consécutive de plusieurs trames correspondant à la même catégorie audiovisuelle. En conséquence, les caractéristiques temporelles d'acquisition et de gestion des données conditionnant l'apprentissage du M-SOM ainsi que le principe de focalisation induit un apprentissage extrêmement rapide des données perçues : les deux sous-réseaux vont être confrontés, durant un certains nombre de trames, à des données similaires et dont le taux d'apprentissage couplé augmente à chaque trame. Ainsi, la création d'une zone dédiée à une nouvelle catégorie — ou le renforcement d'une zone déjà existante — va se faire plus rapidement que dans le cas traditionnel pour lequel des vecteurs d'une matrice entière (déjà connue) sont pris au hasard, afin justement d'éviter le sur-apprentissage et la propagation trop importante des premiers exemples appris. Ces résultats ne sont valables que dans notre cas, c'est-à-dire celui d'un apprentissage en ligne ne disposant pas de mémoire des données perçues [17] et devant inclure à chaque trame le denier exemple disponible afin d'affiner au plus vite la connaissance du robot de son environnement.

Il est donc possible, étant données les caractéristiques de notre problème, de réduire considérablement le nombre d'itérations nécessaires à un apprentissage performant, en comparaison aux nombres employés dans la littérature et pour l'utilisation traditionnelle d'un SOM. Nous validons ainsi l'**Hyp. 7**.

6.5.5 Comportement du M-SOM

Afin d'observer l'évolution de l'apprentissage du M-SOM, nous avons mené une simulation sur $T = 500$ itérations et avons enregistré régulièrement, durant la simulation, les matrices des vecteurs de poids des sous-réseaux audio et visuel.

La **Fig. 6.15** illustre l'évolution de la U-Matrix modifiée (distance entre les vecteurs de poids de chaque neurone des deux sous-réseaux). Premièrement, nous observons la stabilisation progressive du réseau, comportement attendu et caractéristique des algorithmes d'apprentissage de type SOM. Nous pouvons distinguer deux périodes globales :

1. $t = 1$ à $t = 200$: le réseau intègre rapidement toutes les nouvelles données audiovisuelles perçues et s'étend,
2. $t = 200$ à $t = 500$ le réseau est stable.

Deuxièmement, nous observons un comportement essentiel du module MFI : le réseau ne s'auto-organise pas en entier. Nous voyons en effet qu'à partir de $t = 200$

17. La notion de mémoire est ici à différencier de celle dont nous dotons le robot grâce à son apprentissage des environnements. La mémoire des données perçues correspondrait à garder en registre tous les vecteurs de probabilités issus des experts d'identification lors de l'expérience du robot.

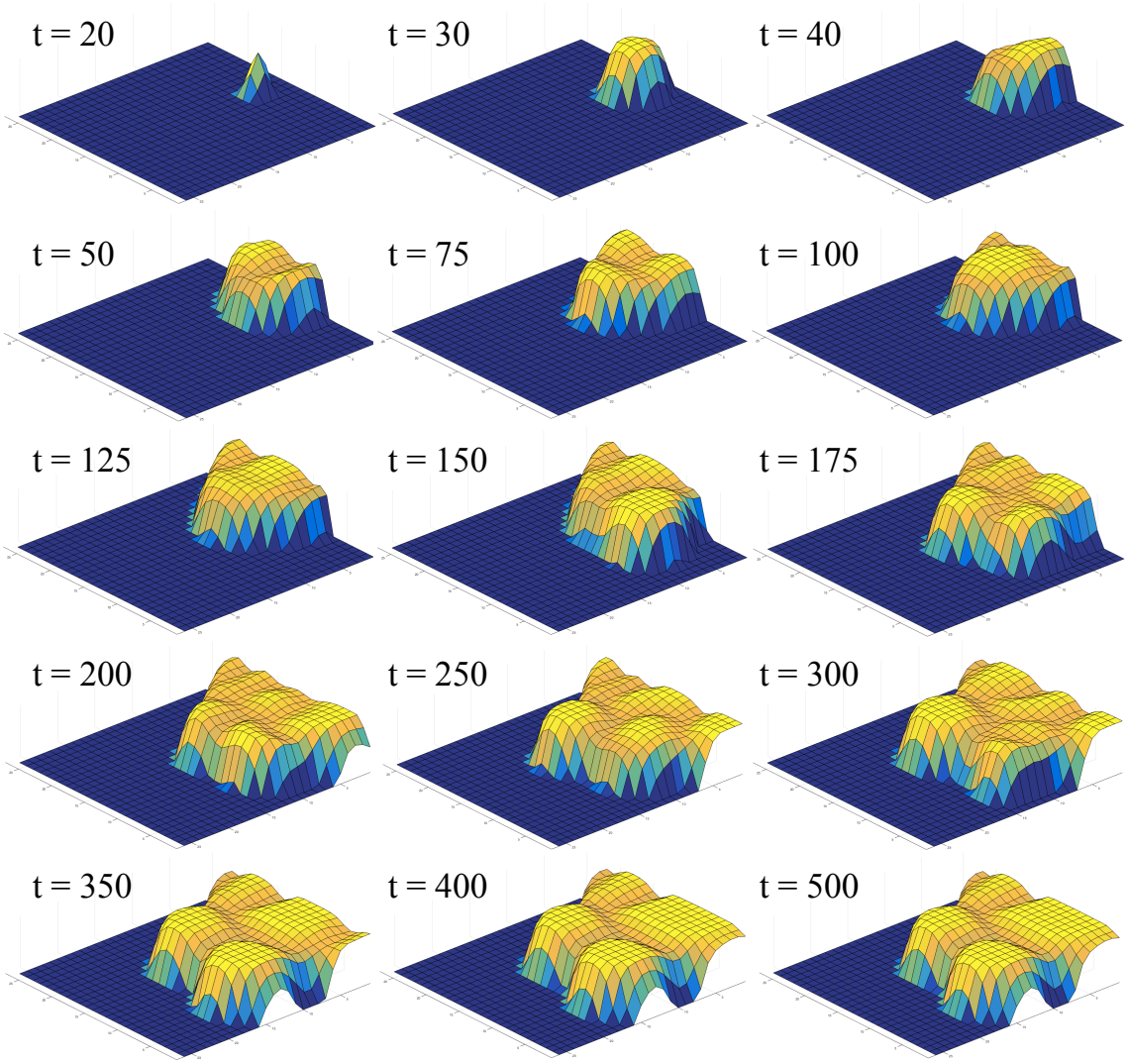


FIGURE 6.15 – EVOLUTION DU M-SOM AU COURS DE L'EXPLORATION — U-Matrix modifiées (cf. **Sec. 6.2.4**) au cours de l'exploration de l'environnement simulé. Pour des raisons de clarté de la représentation, l'opposé des distances a été calculé : (*bleu*) distances élevées, (*jaune*) distances faibles.

environ, le réseau devient très stable (cette rapidité est évidemment très dépendante de la complexité de l'environnement) et que toute une partie du réseau reste insensible à la réorganisation du réseau. Cette caractéristique est essentielle en cela qu'elle permet de laisser au M-SOM la capacité de créer de nouvelles catégories si de nouveaux exemples audiovisuels apparaissent à l'avenir.

Nous avions formalisé le critère $q(\mathcal{C}^{a,v})$ à la section traitant de la convergence du M-SOM. En effet, les observations que nous pouvons faire sur le M-SOM sont principalement dues à cette mesure de la confiance que le module MFI porte en son apprentissage. Sachant que lorsque cette confiance, par catégorie, est suffisamment élevée, les ordres moteurs vers les sources appartenant à cette catégorie sont inhibés, le module MFI n'a pas plus accès aux données complètes et ne peux plus effectuer d'apprentissage. La stabilisation du réseau correspond donc également à la stabilisation des critères $q(\mathcal{C}^{a,v})$ par rapport à K_q (cf. la **Sec. 6.5.2** pour l'étude de l'impact

de la valeur de K_q sur le comportement du module MFI).

6.5.6 Différents environnements

Le module MFI cherche à apprendre le mieux possible les objets audiovisuels présents dans l'environnement en cours d'exploration par le robot. Un des intérêts de cet apprentissage est de pouvoir réutiliser les connaissances acquises dans de nouveaux environnements, intérêt dont nous avons tiré profit pour le module DW. Tous les résultats concernant le module MFI présentés jusqu'à présent ont été obtenus par exploration d'un seul environnement inconnu. Nous allons maintenant étudier le comportement du module MFI lors de l'observation successive de plusieurs environnements. Afin de tester la façon dont le module MFI améliore l'analyse d'un nouvel environnement inconnu sur la base de son expérience passée, nous avons créé le scénario évolutif suivant :

1. Le robot explore un tout premier environnement $e^{(1)}$. Il n'a aucune connaissance sur les catégories audiovisuelles ;
2. Le robot est situé dans un nouvel environnement $e^{(2)}$ totalement différent du premier ;
3. Le robot est situé dans un troisième environnement $e^{(3)}$, en tout point similaire au premier ;
4. Le robot est situé dans un quatrième environnement $e^{(4)}$ dans lequel certains des objets audiovisuels ont déjà été observés dans les deux premiers environnements, d'autres sont nouveaux.

Dans ce scénario composé de quatre environnements à explorer (cf. **Tab. 6.10**), nous nous attendons à observer une inhibition complète des mouvements de tête lorsque le robot explorera le troisième environnement : en effet, sur la base des connaissances aquises sur le premier environnement — et s'il a suffisamment bien appris — aucun mouvement de tête ne devrait être requis pour explorer le dernier environnement.

Conditions de test 6.5.6 [18]						
$e^{(i)}$	$n_{\mathcal{S}}$	n_{sim}^{max}	T	Catégories présentes [19]	K_q	ε_p
1	5	3	500	5, 11, 15, 18	0.8	0.3
2	3	1	500	3, 1	0.8	0.3
3	5	3	500	5, 11, 15, 18	0.8	0.3
4	8	3	500	1, 3, 9, 11, 18, 28, 41	0.8	0.3

TABLE 6.10 – Caractéristiques des 4 environnements générés pour étudier la capacité du module MFI à utiliser les connaissances apprises pour l'exploration de nouveaux environnements inconnus.

La **Fig. 6.16** illustre le nombre de mouvements de tête générés par le module MFI et par le robot naïf dans chacun des quatre environnements. Dans les environnements $e^{(1)}$ et $e^{(2)}$, le nombre de mouvements de tête générés est cohérent avec les résultats obtenus précédemments : le module MFI génère un certain nombre de

18. Se référer à la **Sec. 4.3** pour l'explication de ces notations.
19. Se référer au **Tab. 4.1** pour la liste des catégories audiovisuelles.

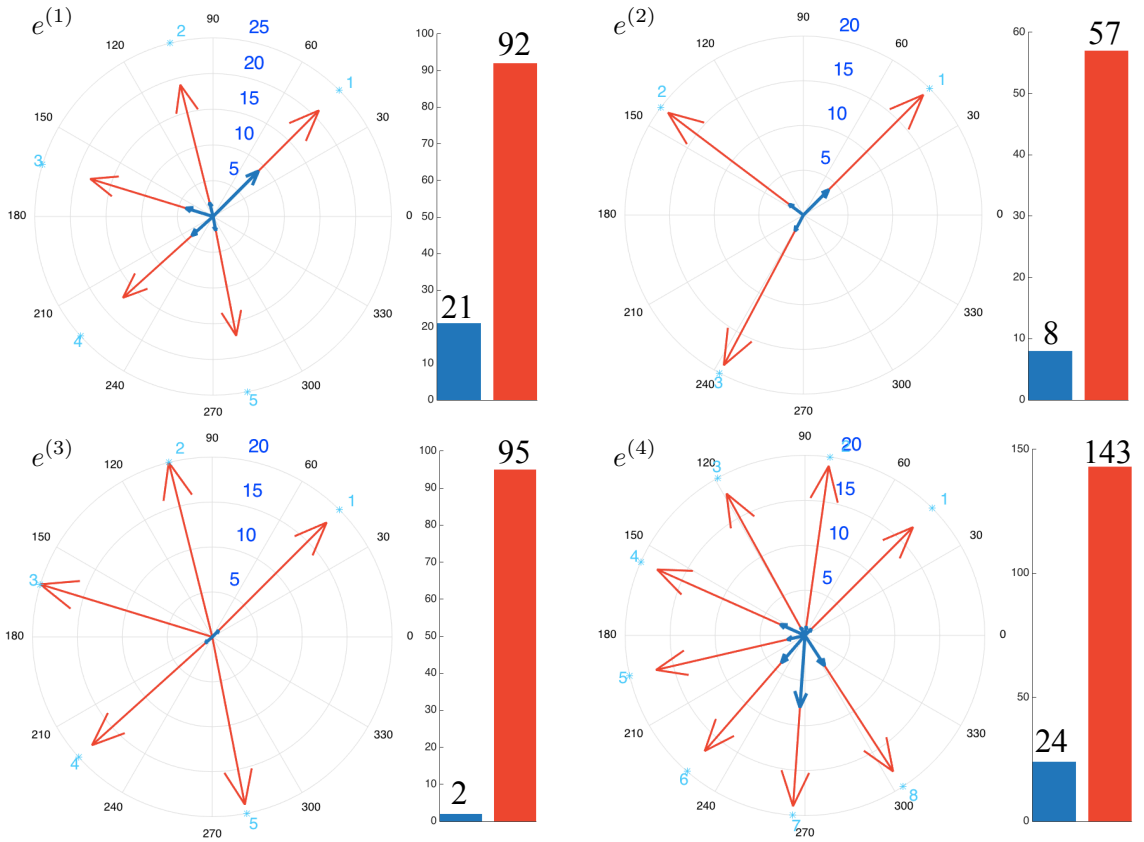


FIGURE 6.16 – NOMBRE DE MOUVEMENTS DE TÊTE ET TRANSMISSION DES CONNAISSANCES — Illustration du nombre de mouvements de tête générés par (*bleu*) le module MFI, (*rouge*) le robot naïf motivé par le principe d'attraction par la Nouveauté. Chaque flèche pointe vers la position d'une source audiovisuelle et leur longueur représente le nombre de mouvements générés vers chacune des sources. Les histogrammes sont la somme de tous les mouvements, indifféremment de la source considérée. Chaque figure représente un environnement différent. Ces environnements ont été explorés successivement, permettant ainsi d'observer le mécanisme de la transmission des connaissances effectué par le module MFI.

---

mouvements de tête afin d'apprendre l'environnement, nombre systématiquement inférieurs à ceux générés par le robot naïf. L'environnement $e^{(3)}$ est celui qui nous intéresse puisqu'il correspond exactement, du point de vue des sources audiovisuelles présentes, non de leur décours temporel, à $e^{(1)}$. Nous observons, dans cet environnement, une inhibition quasi complète des mouvements de tête. Le module MFI se sert des connaissances aquises durant ses précédentes explorations et peut ainsi, lorsqu'il est situé dans un nouvel environnement inconnu, catégoriser correctement les données audiovisuelles perçues de façon quasi instantanée. L'environnement $e^{(4)}$ nous sert de preuve que le M-SOM peut continuer d'apprendre de nouvelles données audiovisuelles et ce grâce au fait que le réseau ne converge jamais totalement.

La **Fig. 6.17** présente justement les taux de bonne classification $\bar{\mathbf{\Gamma}}_{\text{MFI}}$ pour chaque environnement (nous n'avons ici pas indiqué les taux du robot naïf, par souci de clarté, mais avons indiqué le taux d'erreur $\varepsilon_{\mathcal{P}}$). Tous les taux $\bar{\mathbf{\Gamma}}_{\text{MFI}}[T]$ convergent vers 1 mais ce qui nous intéresse particulièrement est la courbe correspondant à

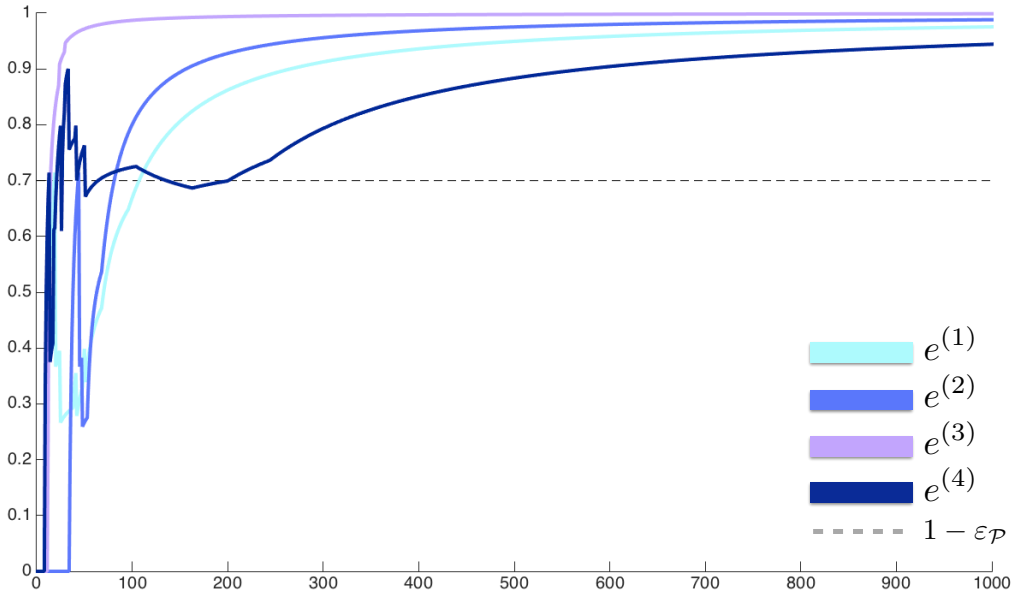


FIGURE 6.17 – TAUX DE CLASSIFICATION DANS DIFFÉRENTS ENVIRONNEMENTS — Chaque courbe représente le taux de classification $\bar{\Gamma}^{a}_{\text{MFI}}$ dans chacun des quatre environnements simulés.

l'environnement $e^{(3)}$, environnement en tout point identique à $e^{(1)}$, du point de vue de leur contenu audiovisuel. Pour rappel, seulement deux mouvements de tête ont été générés par le module MFI dans $e^{(3)}$: la totalité des catégorisations effectuées dans cet environnement ont donc été issues d'inférences. Nous observons que le taux de classification dans $e^{(3)}$ converge beaucoup plus rapidement que les autres vers 1 et est également moins sensible à l'apparition de sources dans l'environnement, comme constaté parfois dans les autres environnements (diminution soudaine et brutale du taux de classification).

La **Fig. 6.18** présente les U-Matrix modifiées du M-SOM en fonction des environnements explorés (cf. **Sec. 6.2.4**). Chaque cellule représente un nœud et la couleur correspond à la distance entre un nœud et ses voisins directs (cf. fonction de voisinage) : plus la couleur tend vers le jaune, plus les nœuds sont proches. La couleur bleue (distances égales) correspond à la partie modifiée du calcul de la U-Matrix et représente ainsi la zone non apprise du M-SOM. Nous utilisons cette représentation afin d'étudier la convergence locale du M-SOM, c'est-à-dire la zone apprise (couleur jaune).

Plusieurs observations sont à faire ici. La première est que l'état du réseau n'a pas changé de l'environnement $e^{(2)}$ à $e^{(3)}$ ce qui est normal puisque quasiment aucune étape d'apprentissage n'a été effectuée. La deuxième est qu'entre les environnements $e^{(1)}$, $e^{(2)}$ et $e^{(4)}$, nous observons que le M-SOM s'est étendu, profitant de l'espace disponible afin d'apprendre les nouvelles catégories audiovisuelles perçues lors de l'exploration. Enfin, nous observons la capacité d'adaptation du M-SOM aux environnements qu'il explore tout en conservant sa capacité à inférer et catégoriser de façon performante les anciennes catégories apprises (cf. **Fig. 6.17**).

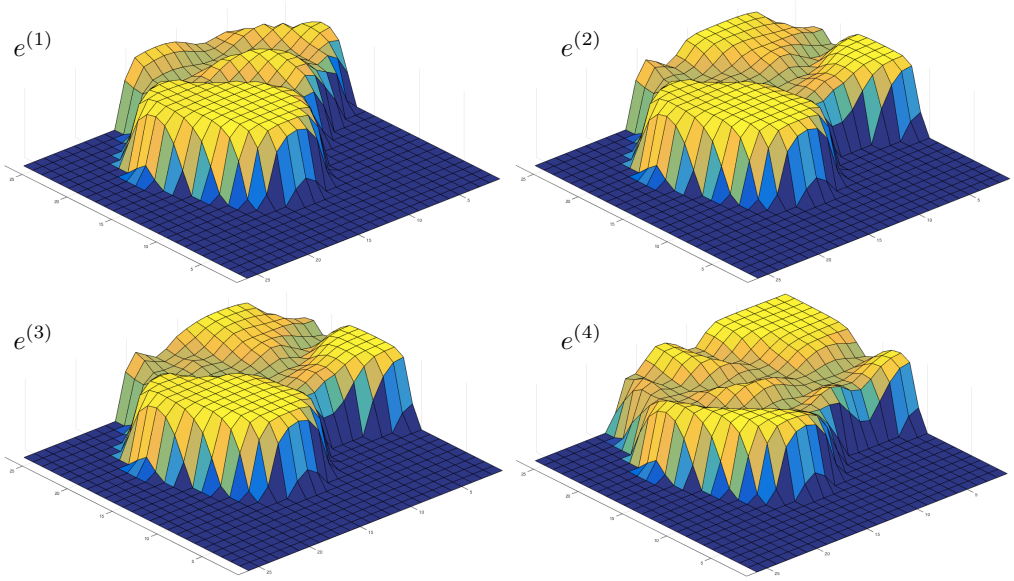


FIGURE 6.18 – CONVERGENCE DU M-SOM DANS DIFFÉRENTS ENVIRONNEMENTS — U-Matrix modifiées (cf. **Sec. 6.2.4**) du M-SOM pour chaque environnement.

6.5.7 Discussion

Cette section a présenté les différentes évaluations du module MFI telles que les performances de classification audiovisuelle, les capacités d'inférence de données manquantes ou la capacité à utiliser les connaissances apprises dans le passé pour analyser plus efficacement de nouveaux environnements inconnus et être capable de diminuer parfois considérablement le nombre de mouvements de tête. Une des limites les plus fortes du module MFI se présente dans le cas suivant :

1. Le robot apprend dans un premier environnement, dans lequel, entre autres, la catégorie audiovisuelle *male speech* est présente. A la fin de l'exploration de cet environnement, le module MFI est confiant dans sa capacité à inférer une modalité manquante.
2. Le robot explore un nouvel environnement dans lequel une source de catégorie *female speech* est présente derrière le robot.
3. Ne percevant que l'information audio, le label *speech*, le module MFI va effectuer une tentative d'inférence de la donnée visuelle.
4. Cette inférence va aboutir à la catégorie *male speech*, seule catégorie contenant le label *speech* qu'il connait, jusqu'à présent.
5. Faisant confiance en son inférence de cette catégorie, il va conclure que la source présente dans ce nouvel environnement appartient à la catégorie *male speech*.

Nous avons ici un problème qui, bien qu'étant une limite du module MFI, est causé par la précision des experts d'identification. En effet, si un classifieur ne peut pas distinguer une voix de femme d'une voix d'homme, il n'y aucune chance pour que le modèle puisse, à partir de cette donnée, inférer le bon label visuel. Cela correspondrait tout simplement à créer de la connaissance là où il n'y en a pas. Ce problème provient du fait qu'à une catégorie audio peuvent être liées plusieurs catégories vi-

suelles (et vice-versa).

Or cette situation, nous la vivons tous les jours. Nos « classifieurs » à nous sont également sujets à imprécisions. Prenons l'exemple d'un non musicien qui écoute deux trompettistes jouer : il est probable qu'il ne distingue pas les différences de sonorités des deux instruments. En revanche, pour une oreille entraînée, ces deux trompettistes jouent sur deux instruments complètement différents. La pratique de la musique aura entraîné un raffinement de la capacité du musicien à analyser des sons : là où le non-musicien considèrera que les deux trompettes appartiennent à la même catégorie, le musicien pourra les assigner à deux catégories dfférentes, voire même inférer l'interprète, la note, la salle dans laquelle a été enregistrée la piste etc.

Ainsi, cette « limite » apparaissant avec le problème d'appariement un à un des classifieurs n'en est pas vraiment une, du point de vue du module MFI. Elle provient de l'incapacité des experts de distinguer deux types d'informations entre eux. Cependant, une piste d'amélioration du module MFI a été envisagée, permettant de prendre en compte cette donnée. Cette piste sera détaillée au dernier chapitre, celui concernant le travail futur.

6.6 Conclusion du Chapitre

Le module *Multimodal Fusion & Inference* a été développé à la suite du module DW répondant au besoin de celui-ci d'avoir accès à la catégorie audiovisuelle d'un objet afin de calculer sa Congruence à l'environnement en cours d'exploration : notamment, si l'objet est défini comme *incongru*, un mouvement de tête sera alors généré vers lui, formalisation d'une forme de réaction attentionnelle. Mais lorsque l'objet est situé derrière le robot, un mouvement de tête est dans un premier temps nécessaire afin d'accéder à la modalité visuelle pour ensuite déterminer si un mouvement de tête dans sa direction sera généré ou non. Cette situation absurde a motivé le développement du module module MFI qui a eu pour but de faire émerger une représentation multimodale de l'environnement, environnement défini par les objets qui le composent.

Le module MFI peut être compris comme une sorte de mémoire multimodale créée durant l'exploration. En effet, le module va motiver l'exploration du monde afin d'aquérir toute l'information dont il a besoin pour créer une représentation interne de celui-ci robuste et stable. Sur cette base, le module sera capable, à partir d'une seule modalité, de retrouver celle manquante et ainsi pouvoir fournir au module DW une information multimodale quelle que soit l'information qu'il a à disposition.

La conception et le développement du module MFI ont été soumis aux mêmes contraintes que ceux du module DW, à savoir une construction d'une représentation de l'environnement basée le plus possible sur l'expérience du robot. Mis à part la connaissance préalable de l'ensemble des catégories audio et visuelles disponibles (*via* l'utilisation des experts d'identification), le système ne se base effectivement que sur les données issues de ces experts lors de son exploration. Pour construire cette représentation, il est nécessaire d'apprendre les données perçues par le robot. Nous avons choisi d'utiliser une carte auto-organisatrice afin de réaliser cet apprentissage, algorithme bien connu et intensivement étudié depuis une quarantaine d'années, nous

permettant de le maîtriser complètement. Mais cet algorithme, tel quel, a des limites qui nous ont amené à proposer une série de modifications substantielles, formalisées par la *Multimodal Self-Organizing Map*. Cette contribution est une extension de l'algorithme SOM traditionnel et permet d'incluer, conceptuellement, tout vecteur de caractéristiques définissant un objet. Dans notre cas, nous avons utilisé l'identification audio et visuelle seulement (même si une extension aux données de localisation a été proposée, détaillée plus tard). A partir de ces données, le module MFI ajoute une nouvelle analyse de l'environnement exprimée par l'ensemble des catégories audiovisuelles qui ont été perçues.

Nous précisons d'ailleurs ici un point important : le module MFI (comme le module DW), ne traite pas les *labels* en eux-mêmes mais leur association. Les classes audio et visuelles doivent être comprises comme « information d'une modalité a » et « information d'une modalité b ». Le but du module MFI est de pouvoir apporter une couche d'analyse supplémentaire des informations disponibles entrant dans la définition d'un objet. Cette définition est, encore une fois, largement extensible et le module MFI (ainsi que le module DW) tel qu'il a été conçu peut être utilisé avec plus de deux modalités.

D'autre part, nous précisons également que le M-SOM n'est pas suffisant pour expliquer tout le comportement du module MFI. L'architecture computationnelle entière du module concourt à l'élaboration d'une représentation interne de l'environnement performante, en terme d'objets audiovisuels qui le composent, permettant (i) de corriger les erreurs des experts d'identification (jusqu'à des taux très hauts), (ii) d'inférer des données manquantes, (iii) de juger de la qualité de son apprentissage et (iv) de générer des ordres moteurs en fonction justement de cette qualité d'apprentissage. Le fonctionnement du module MFI peut ainsi être une nouvelle fois rapproché de celui du colliculus supérieur (cf. **Sec. 2.3.1.3**) en cela que sur la base d'entrées audiovisuelles, il est susceptible de générer un ordre moteur. Il peut également être vu comme un moteur motivationnel de type *réduction de l'incertitude* (cf. **Sec. 2.1.3**).

Nous avons désormais effectué la description et la validation des deux modules du modèle HTM, pris séparément. Depuis le début de la description du modèle, nous avons indiqué que le module MFI a un rôle de support du module DW en cela qu'il lui permet d'avoir accès à des données « propres ». Nous allons ainsi maintenant détailler la mise en commun de ces deux modules, les valider en simulation puis présenter les premiers résultats obtenus sur le vrai robot.

Chapitre 7

Combinaison des modules et Intégration sur le robot

Maintenant que les deux modules constitutifs du modèle HTM ont été détaillés, formalisés et testés séparément, il est temps de les combiner afin d'observer la conjonction de leur action. Ce chapitre est donc dédié à la validation du modèle HTM entier incluant l'effet du module DW dans la génération des mouvements de tête vers des sources audiovisuelles détectées comme incongrues à l'environnement en cours d'exploration, d'une part, et l'effet du module MFI dans l'analyse préalable des informations issues des experts aboutissant également à la génération de mouvements de tête.

La Sec. 7.1 sera dédiée à la description de la façon dont ces deux modules sont connectés, notamment du point de vue de la décision du module à l'origine des mouvements de tête. L'étude du comportement du robot soumis aux deux modules, notamment en mesurant le nombre de mouvements de tête et la qualité de la classification audiovisuelle, sera également menée.

La Sec. 7.2 quant à elle décrira l'implémentation du modèle en tant que Knowledge Source au sein du Blackboard de Two!Ears. Une série d'évaluations menées sur la plateforme robotique de l'ISIR, Odi, sera également effectuée.

7.1 Regroupement des deux modules

Cette section porte sur le regroupement du module DW et du module MFI, regroupement ayant entraîné de légères adaptations, notamment du fait de la capacité de ces deux modules de générer indépendamment des mouvements de tête vers une source d'« intérêt ». Ensuite, l'impact du module MFI sur le module DW, du point de vue de sa dynamique temporelle, sera étudié. Enfin, une section sera dédiée à l'évaluation et à la validation du modèle HTM entier, en conditions simulées.

7.1.1 Ordres moteurs

Les deux modules, de façon indépendante, sont capables de générer des mouvements de tête, comme nous l'avons décrit et observé plus haut. Le module DW est capable de déterminer si un objet nécessite l'attention du robot, grâce au calcul de sa Congruence, tandis que le module MFI est capable de juger si sa connaissance de la catégorie audiovisuelle à laquelle appartient l'objet est suffisamment bonne pour ne pas requérir l'aquisition d'informations supplémentaires à son sujet. Ainsi, à un temps t, il est possible d'avoir deux ordres moteurs contradictoires et il est ainsi nécessaire de prendre une décision sur le module ayant la priorité.

Selon la même approche que celle employée pour les deux modules séparément (selon le modèle GPR décrit à la **Sec. 2.1.2.2**), nous allons déterminer quel module à la priorité en considérant que l'ordre moteur généré par chaque module est un canal d'information ayant une activité spécifique. Pour rappel, le modèle GPR tente de formaliser le rôle de la boucle ganglions de la base — thalamus — cortex dans la sélection des actions motrices, indispensable lorsque deux actions contradictoires sont possibles, comme tourner à gauche ou à droite, par exemple. Le modèle GPR considère que chaque action motrice est codée par un canal d'information possédant une certaine activité. Tous les canaux sont inhibés par défaut et celui ayant l'activité la plus faible se verra désinhibé.

Un point important ici est celui du but de l'ordre moteur généré par chacun des modules : le mouvement requis par le module DW peut être vu comme une réaction attentionnelle tandis que celui requis par le module MFI est une réaction à une mauvaise connaissance de l'environnement. Si le module DW prend une décision sur la base d'une mauvaise catégorisation audiovisuelle de la part du module MFI, la réaction motrice sera erronnée. Ainsi, nous considérons que le besoin du module MFI d'affiner sa capacité à comprendre l'environnement, formalisé par le critère q, est plus important que la réaction attentionnelle déclenchée par le module DW. Nous allons donc redéfinir l'expression de l'activité du module DW afin de prendre en compte cette donnée.

Soit le symbole de Kronecker $\delta^{(k)}[t]$ défini comme :

$$\delta^{(k)}(x) = \begin{cases} 1 \text{ si } x >= 1, \\ 0 \text{ sinon.} \end{cases} \tag{7.1}$$

Nous redéfinissons alors l'activité du module DW en tenant compte de sa modulation par l'activité du module MFI selon :

$$\tau_{\text{DW}}[t] = \tau_{\text{MFI}}[t] - \tau_{\text{DW}}[t] \times \delta^{(k)}(\tau_{\text{MFI}}[t]) \tag{7.2}$$

avec $\tau_{\text{MFI}}[t]$ défini par l'**Eq. 6.10** comme le rapport entre q et K_q. Nous rappelons que si l'activité du module MFI dépasse la valeur seuil de 1, cela signifie que le module fais confiance en l'inférence des catégories de toutes les sources présentes et qu'aucun mouvement de tête ne sera généré. Ainsi, étant donnée l'expresion de l'**Eq. 7.2** :

- l'activité du module DW sera supérieure à celle du module MFI tant que cette dernière sera inférieure à 1,

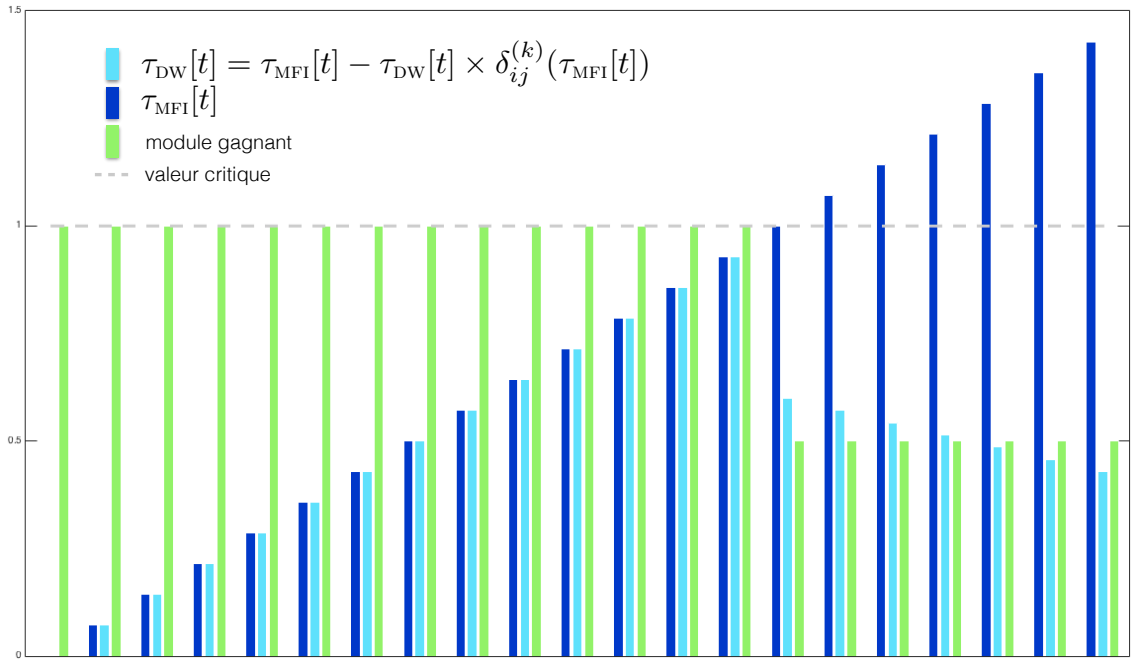


FIGURE 7.1 – ACTIVITÉ COMBINÉE DU DW ET DU MFI — Illustration de l'impact de l'activité du module MFI sur celle du module DW selon la nouvelle expression de cette dernière. (*bleu foncé*) activité du module MFI et (*bleu clair*) activité du module DW. (*vert*) activité combinée. Une valeur de 1 indique que le module MFI l'emporte, une valeur de 0.5 indique que c'est le module DW. La ligne pointillée indique la valeur à partir de laquelle le module MFI ne génère plus de mouvements de tête, moment à partir duquel le module DW prend le dessus.

- l'activité du module DW sera inférieure à celle du module MFI dès que cette dernière dépassera 1.

Suivant la règle de détermination de l'action motrice à promouvoir par sélection du canal à l'activité la plus faible, la modification de l'activité du module DW que nous venons d'effectuer permet de laisser le module MFI prendre le dessus lorsque celui-ci a besoin. A l'opposé, lorsque module MFI ne génère plus de mouvements de tête, le module DW peut prendre le relais.

7.1.2 Impact du MFI sur le DW

Pour les mêmes raisons qui nous ont amené à modifier l'expression de l'activité du module DW dans le calcul des commandes motrices combinées, nous avons modifié les conditions d'exécution du module DW en fonction du module MFI. Ainsi, le calcul de la Congruence d'un objet o_j ne se fera que lorsque la confiance du module MFI en sa connaissance de la catégorie audiovisuelle à laquelle cet objet appartient sera suffisamment grande. Une nouvelle fois, le critère $q(\mathcal{C}^{a,v})$ va nous permettre de formaliser ce comportement : tant que $q(\mathcal{C}^{a,v}) < K_q$, le module DW n'effectuera aucun calcul de Congruence. Ce n'est que lorsque le système est sûr que les données qui lui sont envoyées peuvent être considérées comme correctes et stables que le module DW commencera à analyser l'objet en question. Ce comportement aura

pour conséquence, lors de l'exploration d'un environnement, une dynamique en trois phases :

1. une première phase durant laquelle le module MFI est le seul module responsable des mouvements de tête ;
2. une deuxième phase durant laquelle certaines catégories sont bien apprises et pour lesquelles le module DW est susceptible de générer un ordre moteur, tandis que d'autres sont encore mal apprises ;
3. une dernière phase durant laquelle toutes les catégories audiovisuelles ont été bien apprises par le module MFI et le module DW est le seul module pouvant générer des mouvements de tête.

A noter que les mouvements de tête générés par le module DW serviront également le module MFI : l'acquisition de données audio et visuelles sur un objet pourront être utilisées par le module MFI pour continuer l'apprentissage. Cependant, le fait que le module DW n'intègre pas de facteur temporel dans son calcul de Congruence pourrait aboutir à un sur-apprentissage d'une catégorie audiovisuelle par le M-SOM. En effet, si le module DW génère des ordres moteurs sur une catégorie en particulier et pendant un certain temps, le nombre d'itérations d'apprentissage que le M-SOM va effectuer pour cette catégorie va devenir, en proportion, beaucoup plus important que pour les autres catégories. La propagation plus importante de l'apprentissage de ces données — due au fait que nous n'utilisons pas le SOM sur une matrice entière de données au sein de laquelle un vecteur est pris au hasard à chaque étape d'apprentissage — résulterait dans l'attraction conséquente de la zone « sur-apprise » pour d'autres vecteurs différents. Afin d'éviter ce cas critique mais (i) de conserver le principe selon lequel le module MFI détermine lorsqu'il doit continuer d'apprendre ou non et (ii) de profiter malgré tout de l'apport de nouvelles données, nous avons plafonné l'itération d'apprentissage de l'objet considéré lorsque les données audiovisuelles acquises ont été le résultat d'un mouvement ordonné par le module DW et non par le module MFI. Ainsi, et dans ce cas seulement, le système impose $n_{it}[t] = 5$ (d'après la **Eq. 6.6**), soit $N_{it}/2$.

D'autre part, nous avons vu à la **Sec. 5.3.3** que le module DW est capable d'utiliser les connaissances qu'il a apprises lors de l'exploration de précédents environnements pour analyser plus rapidement un nouvel environnement inconnu dans lequel il se trouve. Or cette transmission de connaissances, formalisée par la **Déf. 13** et son corollaire, se base sur une base de données des catégories audiovisuelles détectées jusqu'à présent. Cette base de données est directement issue du module MFI et de son analyse du M-SOM. En effet, toutes les catégories que le modèle HTM traite sont issues de l'auto-organisation du M-SOM. Ainsi, nous pouvons considérer qu'il existe une instance du module DW par type d'environnement (caractérisé par les catégories audiovisuelles qui le composent) mais qu'il n'existe qu'une seule instance du module MFI et donc qu'un seul M-SOM en charge de l'apprentissage de l'espace audiovisuel entier.

Afin de valider ce comportement et d'étudier l'impact du rassemblement des deux modules sur les performances du système entier, nous avons effectué une série de tests, similaires à ceux effectués sur les modules séparément. La prochaine section présente les résultats obtenus.

7.1.3 Résultats

De façon similaire aux évaluations menées sur les deux modules séparément, nous avons créé divers environnements de test permettant d'observer les performances de fusion et de classification du modèle entier ainsi que l'impact sur les mouvements de tête, toujours en comparaison du robot naïf. Cependant, nous avons vu à la section précédente que la conjonction des deux modules consiste en quelque sorte en leur juxtaposition : pour un objet appartenant à une catégorie audiovisuelle donnée, lorsque le module MFI juge que sa connaissance de cette catégorie est suffisante, il laisse la place au module DW. Ainsi, l'ensemble des résultats obtenus sur les différentes parties des deux modules (comme le critère K_q ou la persistance temporelle) seront retrouvés ici. Nous proposons ainsi ici de ne présenter l'évaluation globale du module que selon trois ensembles de conditions de test : en cas unisource, en cas multisource puis dans différents environnements successifs. Afin de mesurer les performances du modèle entier, nous utiliserons les taux de bonne classification utilisés pour l'évaluation du module MFI ainsi que le nombre de mouvements de tête générés par l'ensemble du modèle. Notre point de comparaison sera une nouvelle fois le robot naïf, omniscient ou non. Nous rappelons que le robot naïf est doté (i) d'un système de fusion des données directement en sortie des classifieurs et (ii) doté d'un comportement motivé par une interprétation de la saillance des événements (l'apparition d'une source sera saillante par rapport au contexte).

Le **Tab. 7.1** liste les conditions de test dans des cas unisource et multisources, de façon similaire à l'évaluation du module MFI effectuée au chapitre précédent.

Conditions de test 7.1.3 [1]						
$e^{(i)}$	$n_{\mathcal{S}}$	n_{sim}^{max}	T	Catégories présentes [2]	K_q	ε_p
Cas unisource						
1	3	1	500	1, 9	0.8	0.3
2	5	1	500	3, 1	0.8	0.3
3	7	1	500	5, 11, 15, 18	0.8	0.3
4	10	1	500	1, 3, 9, 11, 18, 28, 41	0.8	0.3
Cas multisource						
5	3	3	500	1, 9	0.8	0.3
6	5	5	500	3, 1	0.8	0.3
7	7	7	500	5, 11, 15, 18	0.8	0.3
8	10	10	500	1, 9, 18, 21, 39	0.8	0.3

TABLE 7.1 – Caractéristiques des 8 environnements générés en cas unisource et multisource générés pour étudier la combinaison du module DW et du module MFI.

7.1.3.1 Bonne classification

Les résultats des taux de bonne classification pour le cas unisource et le cas multisource sont présentés au **Tab. 7.2**. Nous voyons que l'ajout du module DW

1. Se référer à la **Sec. 4.3** pour l'explication de ces notations.
2. Se référer au **Tab. 4.1** pour la liste des catégories audiovisuelles.

Résultats pour les conditions 7.1.3					
$n_{\mathcal{S}} \mid n_{sim}^{max}$	$\bar{\Gamma}^{a'}_{\text{MFI}}[t=T]$	$\bar{\Gamma}^{a''}_{\text{MFI}}[t=T]$	$\bar{\Gamma}_{\mathfrak{R}_n}[t=T]$	$\bar{\Gamma}'_{\mathfrak{R}_n}[t=T]$	ratio
Cas unisource					
3 | 1	0.955 (0.039)	0.960 (0.039)	0.689 (0.012)	$\bar{\Gamma}_{\mathfrak{R}_n}[t=T] = \bar{\Gamma}'_{\mathfrak{R}_n}[t=T]$	1.389
5 | 1	0.947 (0.049)	0.950 (0.049)	0.682 (0.052)		1.390
7 | 1	0.937 (0.058)	0.942 (0.059)	0.648 (0.054)		1.449
10 | 1	0.937 (0.061)	0.942 (0.060)	0.673 (0.026)		1.396
moyenne	**0.944**	**0.948**	**0.673**		**1.406**
Cas multisource					
3 | 3	0.943 (0.024)	0.968 (0.021)	0.693 (0.008)	0.428 (0.053)	1.378
5 | 5	0.908 (0.027)	0.975 (0.007)	0.692 (0.014)	0.272 (0.023)	1.360
7 | 7	0.845 (0.049)	0.929 (0.043)	0.687 (0.010)	0.211 (0.010)	1.291
10 | 10	0.724 (0.074)	0.849 (0.090)	0.687 (0.004)	0.136 (0.006)	1.144
moyenne	**0.855**	**0.930**	**0.689**	**0.261**	**1.293**

TABLE 7.2 – Taux de bonne classification pour l'ensemble des conditions de simulations présentées au **Tab. 7.1**. Chaque résultat est une moyenne sur les 5 répétitions de chaque conditions (avec l'écart-type entre parenthèse), pour un total de 40 simulations. Les valeurs sont arrondies à la troisième décimale.

n'affecte aucunement les performances de classification (fusion ou inférence). Dans toutes les conditions, le modèle, considéré dans son ensemble, est donc capable de fournir au *Blackboard* de façon robuste une connaissance supplémentaire sur l'environnement en cours d'exploration, du point de vue des objets audiovisuels présents.

7.1.3.2 Mouvements de tête

La **Fig. 7.2** présente l'impact de la combinaison des deux modules sur le nombre de mouvements de tête générés, dans le cas unisource, et la **Fig. 7.3** dans le cas multisource (nous rappelons que tous les nombres indiqués sont des moyennes sur cinq simulations effectuées pour chaque condition de test). Premièrement, nous observons que le nombre de mouvements de tête générés par le module DW est systématiquement inférieur à ceux générés par le module MFI. Nous observons, de plus, que le module MFI prend le pas sur le module DW : plus il y a de mouvements de tête générés par le module MFI, moins il y en a par le module DW. Ce comportement s'explique par la durée de la simulation : en faisant durer la simulation plus longtemps, le module DW prendrait progressivement le pas sur le module MFI, ce dernier ayant tendance à faire converger son apprentissage et inhiber conséquemment ses ordres moteurs, au profit du module DW. Nous reviendrons sur ce point un peu plus tard.

D'autre part, nous observons, dans le cas unisource, que le nombre de mouvements de tête générés par le modèle HTM entier est presque toujours inférieur à ceux générés par le robot naïf, exception faite de la condition n°4 ($n_{\mathcal{S}} = 10$). La différence entre les deux systèmes a tendance diminuer avec la complexité de l'environnement, ici concrétisée par le nombre de sources présentes. Mais encore une fois, il s'agit

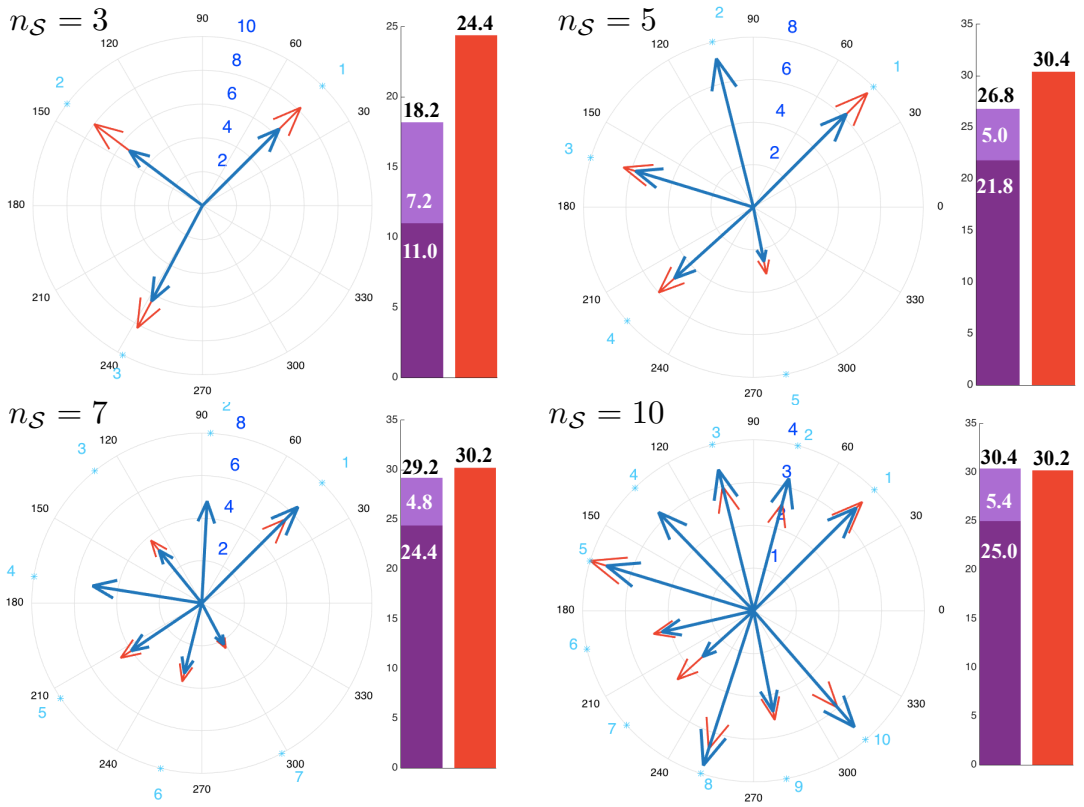


FIGURE 7.2 – NOMBRE DE MOUVEMENTS DE TÊTE GÉNÉRÉS EN CAS UNISOURCE — Mouvements générés (*bleu*) par le modèle HTM entier et (*rouge*) par le robot naïf. Les flèches pointent vers les positions des sources sonores, leur longueur dénotant le nombre de mouvements vers la source concernée. Les histogrammes représentent la somme totale des mouvements générés (*violet foncé*) par le module MFI, (*violet clair*) par le module DW et (*rouge*) par le robot naïf. Les chiffres (*blancs*) correspondent au nombre de mouvements par module, (*noirs*) totaux.

---

d'un effet dû à la durée de la simulation. Afin de s'en assurer, nous avons effectué une simulation sur $T = 2000$ pas de temps avec les paramètres de la condition n°4, dont les résultats sont présentés à la **Fig. 7.4**. Nous voyons là que le nombre de mouvements de tête générés par le modèle est bien inférieur à ceux générés par le robot naïf[3]. De plus, le ratio entre les mouvements générés par les deux modules est plus proche de 1 : le module DW prend petit à petit le pas sur le module MFI.

Revenant à la **Fig. 7.3** et au cas multisource des conditions testées initialement, nous observons un effet inverse de celui constaté en cas unisource : plus l'environnement est complexe, plus la différence entre le modèle HTM et le robot naïf s'accentue.

7.1.3.3 Dynamique des deux modules

Afin d'étudier spécifiquement la dynamique temporelle des deux modules une fois combinés, nous avons effectué deux simulations, une en condition unisource et une en condition multisource, présentées au **Tab. 7.3**.

3. Nous discuterons d'ailleurs à la **Sec. 8.4** l'intérêt de cette comparaison entre le modèle et

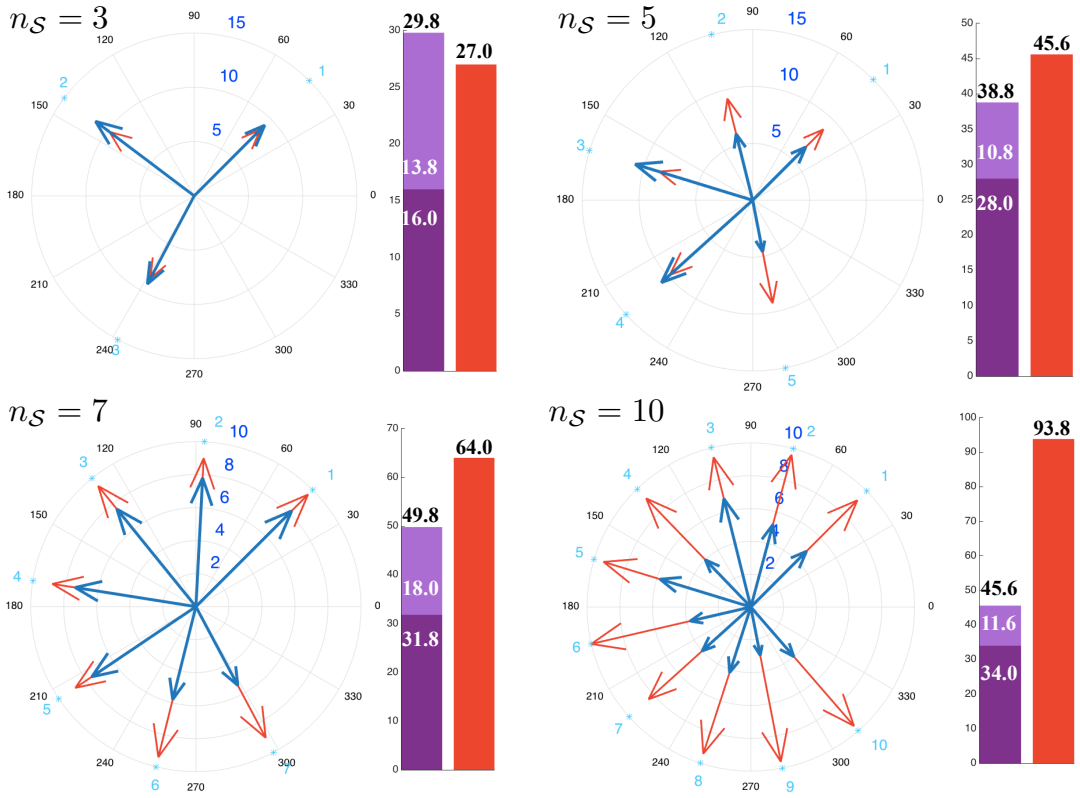


FIGURE 7.3 – NOMBRE DE MOUVEMENTS DE TÊTE GÉNÉRÉS EN CAS MULTI-SOURCE — Mouvements générés (*bleu*) par le modèle HTM entier et (*rouge*) par le robot naïf. Les flèches pointent vers les positions des sources sonores, leur longueur dénotant le nombre de mouvements vers la source concernée. Les histogrammes représentent la somme totale des mouvements générés (*violet foncé*) par le module MFI, (*violet clair*) par le module DW et (*rouge*) par le robot naïf. Les chiffres (*blancs*) correspondent au nombre de mouvements par module, (*noirs*) totaux.

Conditions de test 7.1.3.3 [4]						
n°	$n_{\mathcal{S}}$	n_{sim}^{max}	T	Catégories présentes [5]	K_q	ε_p
1	5	1	500	1, 9, 18	0.7	0.3
2	5	5	1000	1, 9, 18	0.7	0.3

TABLE 7.3 – Caractéristiques de l'environnement généré pour étudier la dynamique temporelle des deux modules une fois combinés.

La **Fig. 7.5** nous permet d'observer (i) les objets focalisés par le robot et (ii) le module responsable du mouvement de tête. Nous voyons clairement ici la façon dont le module MFI est responsable, en début d'exploration, de la génération des mouvements de tête puis, une fois les catégories suffisamment apprises, la façon dont le DW prend le relais devenant le seul module à l'origine des mouvements de tête. Nous retrouvons d'ailleurs bien le comportement en trois phases décrit à la **Sec. 7.1.2** :

le robot naïf.

4. Se référer à la **Sec. 4.3** pour l'explication de ces notations.
5. Se référer au **Tab. 4.1** pour la liste des catégories audiovisuelles.

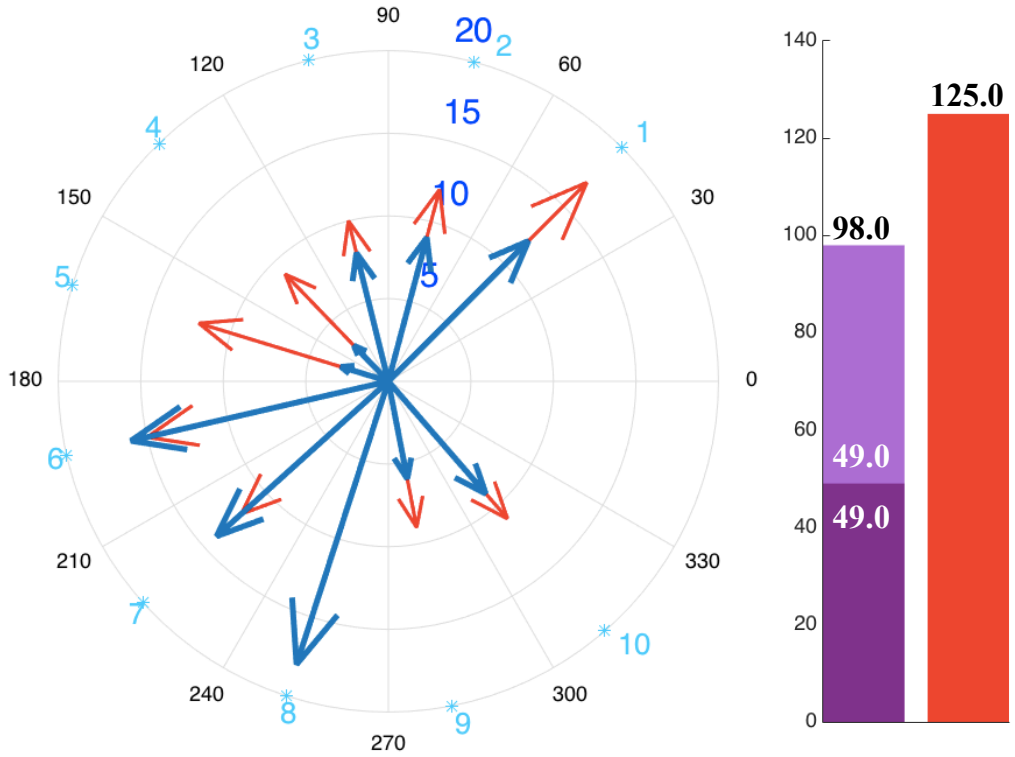


FIGURE 7.4 – NOMBRE DE MOUVEMENTS DE TÊTE GÉNÉRÉS POUR LA CONDITION N°4 — Ici, la durée de simulation a été fixée à $T = 2000$ pas de temps. Mouvements générés (*bleu*) par le modèle HTM entier et (*rouge*) par le robot naïf. Les flèches pointent vers les positions des sources sonores, leur longueur dénotant le nombre de mouvements vers la source concernée. Les histogrammes représentent la somme totale des mouvements générés (*violet foncé*) par le module MFI, (*violet clair*) par le module DW et (*rouge*) par le robot naïf. Les chiffres (*blancs*) correspondent au nombre de mouvements par module, (*noirs*) totaux.

---

1. première phase : le module MFI est le seul module responsable des mouvements de tête ;
2. deuxième phase : certaines catégories sont bien apprises permettant au module DW d'éventuellement générer un ordre moteur, tandis que d'autres sont encore mal apprises ;
3. troisième phase : toutes les catégories audiovisuelles sont désormais bien apprises par le module MFI et le module DW est le seul module pouvant générer des mouvements de tête.

La **Fig. 7.6** (à mettre en rapport avec la **Fig. 7.5**) illustre le ratio entre le nombre de mouvements de tête générés par le DW et ceux générés par le MFI, ratio calculé à chaque pas de temps. Une valeur inférieure à 1 signifie que le MFI a généré plus de mouvements de tête que le DW, une valeur supérieure signifie l'inverse. Nous observons premièrement qu'en début d'exploration, le ratio est largement en faveur du MFI (première phase). Le DW prend graduellement le relais faisant ainsi tendre le ratio vers 1 (deuxième phase). En fin d'exploration, seul le DW génère des mouvements de tête poursuivant l'augmentation du ratio vers 1.

D'ailleurs, la **Fig. 7.7**, issue de la condition n°2, illustre l'évolution du ratio lors

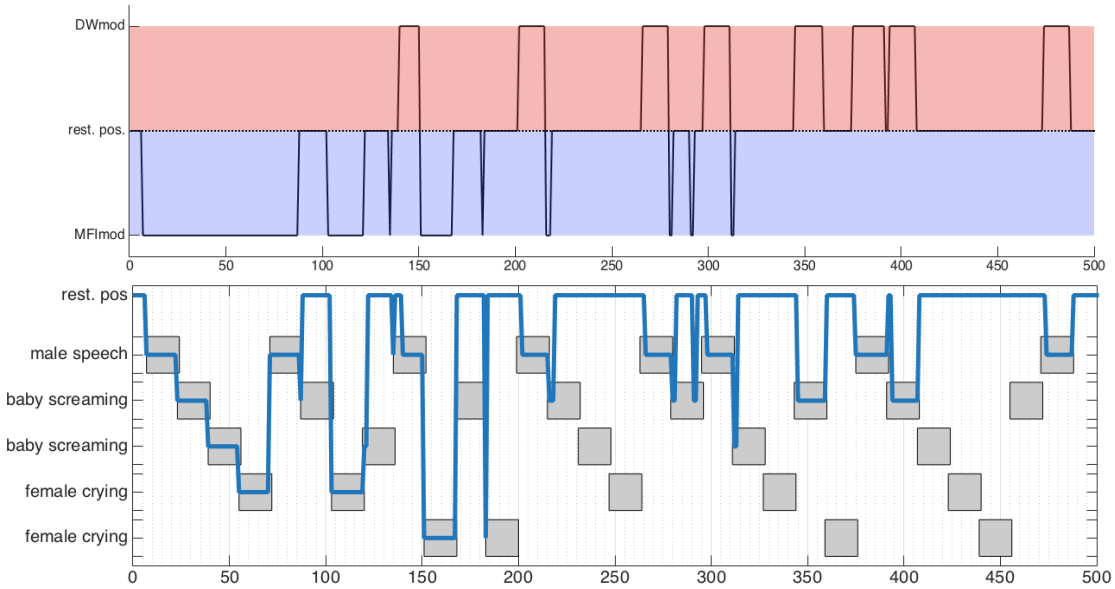


FIGURE 7.5 – DYNAMIQUE DES DEUX MODULES, CAS UNISOURCE — Illustration de l'évolution du module à l'origine des mouvements de tête générés. (*haut*) Mouvements de tête générés (*partie rouge*) par le DW, (*partie bleue*) par le MFI. (*bas*) décours temporel de la simulation et objets focalisés par le système entier.

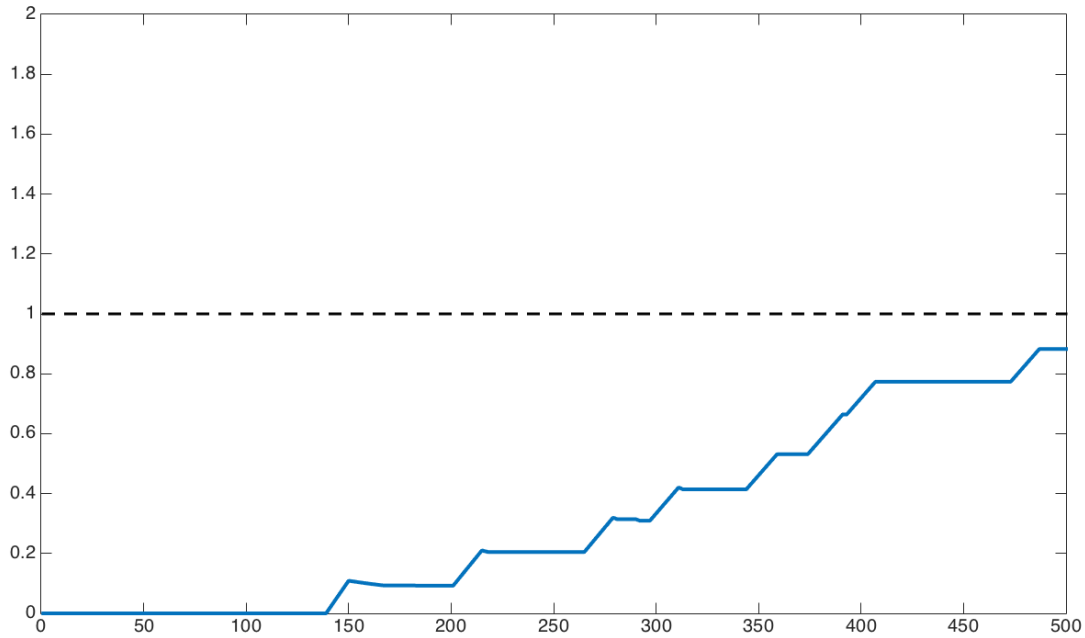


FIGURE 7.6 – RATIO ENTRE LES DEUX MODULES, CAS UNISOURCE — Illustration du ratio entre le nombre de mouvements de tête générés par le MFI et ceux générés par le DW selon une moyenne glissante effectuée à chaque pas de temps de la simulation. Une valeur inférieure à 1 signifie que le MFI a généré plus de mouvements que le DW, une valeur supérieure à 1 signifie l'inverse.

d'une simulation plus longue ($T = 1000$) : à partir de $t = 457$, le ratio dépasse la valeur 1, signifiant qu'à partir de ce temps là le module DW a généré plus de mouvements de tête que le MFI. Jusqu'à la fin de la simulation, le ratio augmente

continuellement puisque l'activité du MFI est complètement inhibée.

7.1.4 Discussion

Nous avons présenté dans cette section la combinaison du module DW et du module MFI constituant ainsi le modèle HTM entier. Notamment, les deux modules requérant indépendemment des mouvements de tête, pour des raisons qui leur sont propres, il a fallu modifier l'expression de l'activité du module DW afin de ne le rendre actif que lorsque le module MFI juge que la catégorie à laquelle un objet considéré est désormais suffisamment apprise par lui. A ce moment-là, le module DW peut commencer à effectuer ses calculs de Congruence aboutissant éventuellement en la génération de mouvements de tête. Cette décision se fera ainsi sur des données « propres » et pertinentes : le module MFI joue ainsi également un rôle dans le comportement attentionnel du robot en cela qu'il assure que les décisons prises par le module DW ne se font pas sur des objets audiovisuels faux entraînant une réaction comportementale non pertinente.

Le modèle entier a été testé en conditions simulées. Les résultats confirment ceux obtenus lors des validations séparées des deux modules constitutifs du modèle HTM, le module DW et le module MFI. Notamment, les taux de bonne classification sont très élevés, quasiment tous supérieurs à ceux obtenus par le robot naïf omniscient et aux capacités non réalistes puisqu'il a accès à toutes les données, tout le temps. Nous avons cependant observé une augmentation globale des mouvements de tête générés par les deux modules combinés. Nous précisons qu'il ne s'agit pas d'un effet

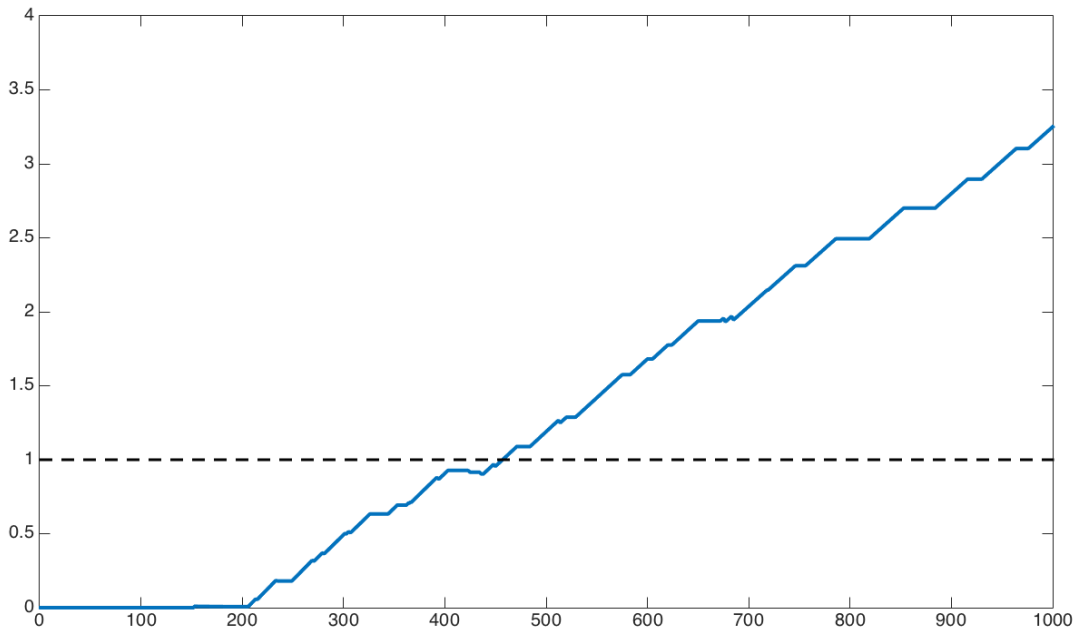


Figure 7.7 – Ratio Entre les Deux Modules, Cas Multisource — Illustration du ratio entre le nombre de mouvements de tête générés par le MFI et ceux générés par le DW selon une moyenne glissante effectuée à chaque pas de temps de la simulation. Une valeur inférieure à 1 signifie que le MFI a généré plus de mouvements que le DW, une valeur supérieure à 1 signifie l'inverse.

d'un des modules sur l'autre mais plutôt du résultat de l'addition des mouvements de tête générés par chacun des modules. D'autre part, le module DW, tel qu'il a été formalisé, tournera sa tête à chaque fois qu'un objet considéré comme incongru apparaîtra dans l'environnement, même s'il s'agit du même objet qui apparaît sans cesse par intermittance. Nous reviendrons, à la **Sec. 8.1**, sur ce point important du module DW et notamment sur les pistes de travail futur le concernant.

La section suivante décrit maintenant l'intégration du modèle HTM au sein du système Two!Ears permettant son utilisation avec la plateforme robotique réelle. Les expériences menées sur Odi, le robot de l'ISIR, et les résultats obtenus seront également décrits.

7.2 Intégration sur le robot et résultats

Tout le travail effectué présenté jusqu'à présent a été évalué en conditions simulées. En effet, jusqu'au mois n°35 du projet, c'est-à-dire jusqu'à début octobre 2016 (deux mois avant la fin officielle du projet et pendant la période de rédaction intense de tous les documents officiels finaux à délivrer pour le mois de novembre 2016), le modèle HTM, implémenté en tant KS, n'a pas eu accès aux robots pleinement fonctionnels pour les besoins du modèle ni à l'intégration stable et robuste de certaines KS indispensables à son fonctionnement. Par exemple, le robot Odi n'a été doté d'un cou fonctionnel (capable de rotation) qu'en septembre 2016, installation ayant été suivie d'une longue phase d'adaptation de certains composants matériels et logiciels pour notre plateforme. Cependant, durant toute cette thèse, un travail continu a été effectué sur Odi, le robot de l'ISIR, en collaboration temporaire avec M. Antonyo Musabini. Finalement, le logiciel Two!Ears ainsi que le modèle HTM ont pu être intégrés et testés sur les deux plateformes robotiques. Cette section consiste donc en l'intégration de la HTMKS au sein de Two!Ears, puis sur la description du travail effectué sur la plateforme robotique Odi, et enfin sur les résultats obtenus en conditions réelles sur Odi.

7.2.1 Implémentation en tant que KS

Le modèle HTM a été implémenté en tant que Knowledge Source et intégré au Blackboard. Cette intégration permet d'assurer la bonne communication avec les autres KS que ce soit pour la récupération de données que pour leur publication. L'entité centrale de l'adaptation du modèle en KS est la HeadTurningModulationKS (HTMKS), mais un ensemble d'autres KS a dû être également implémenté pour adapter le modèle entier au logiciel Two!Ears, comme la **Fig. 7.9** l'illustre. En tant que KS « haut-niveau », c'est-à-dire nécessitant le traitement préalable des signaux audio et visuels et effectuant des analyses plus proches des processus cognitifs que du traitement du signal, la HTMKS est placée parmi les dernières KS au sein du Scheduler : elle est donc une des dernières a être exécutée. Cette KS, comme toutes les autres, a été codée en langage Matlab® selon le paradigme orienté-objet adopté dans le projet.

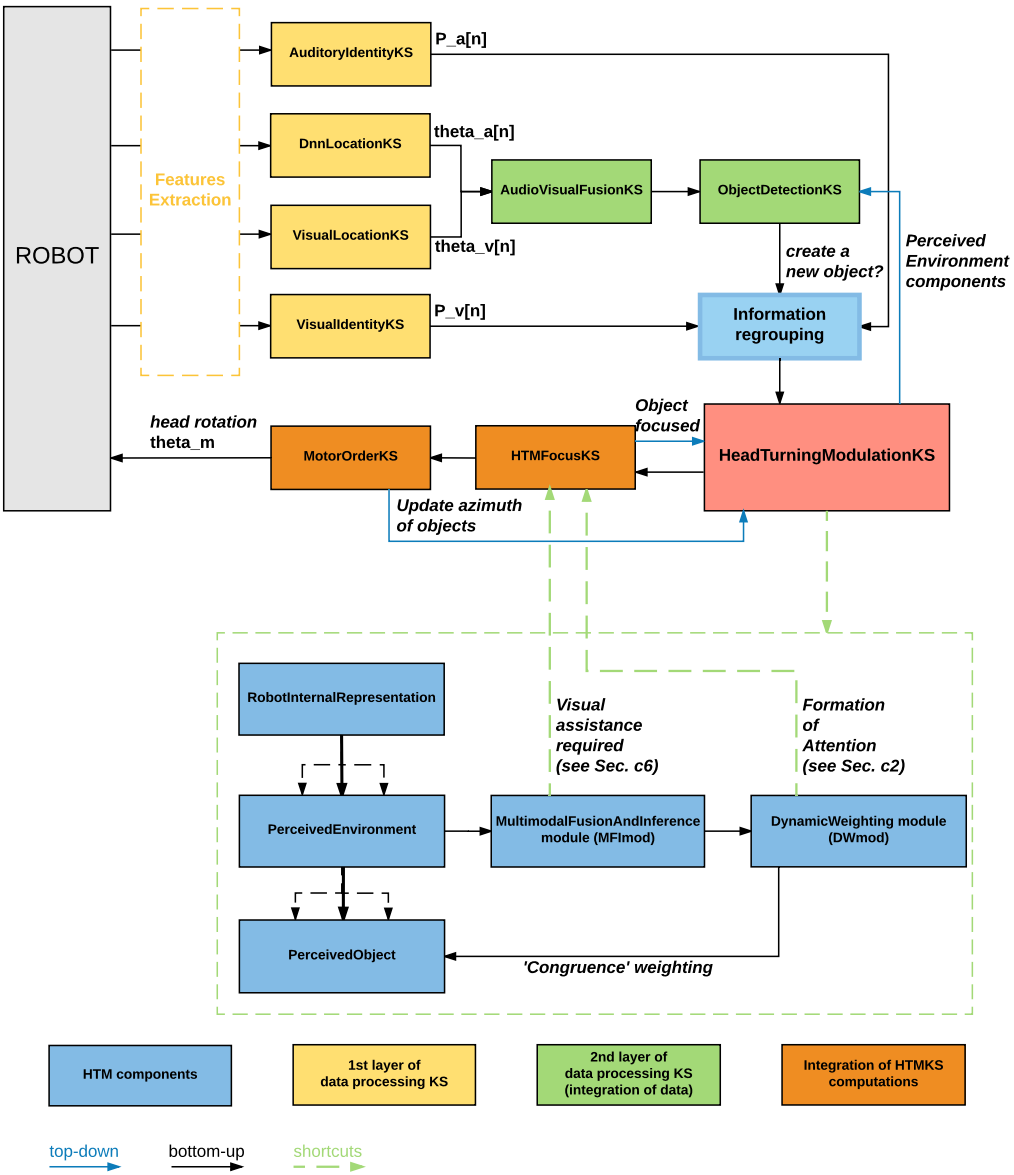


FIGURE 7.8 – ARCHITECTURE DU MODÈLE HTM EN TANT QUE KNOWLEDGE SOURCE — Schéma de la structure du modèle HTM implémenté en tant que KS, ainsi que ses liens avec les autres composants de TWO!EARS. (*rouge*) communication avec le robot *via* la variable ROBOTCONNECT du *Blackboard* ; (*vert*) dépendance de la KS correspondante à l'AUDITORYFRONTEND ; (*bleu*) implémentation en classe MATLAB®, selon le paradigme orienté objet utilisé dans le logiciel TWO!EARS.

La HTMKS est en réalité plus qu'une unique KS : elle est une importante structure contenant plusieurs autres KS et classes (ou sources) (cf. **Fig. 7.8**). La HTMKS peut même être considérée à elle seule comme un système de type *blackboard*. En effet, suivant la définition de DAVID D. CORKILL [226] (cf. **Sec. 3.2.1**) : (i) la HTMKS est constituée d'un ensemble de KS dédiées à une partie de l'analyse des informations reçues ; (ii) ces informations sont reçues sous la forme d'événements provenant du *Blackboard* et *via* la HMTKS justement ; (iii) chacune de ses entités constitutives communique de façon uniformisée et lui transmet directement ses données. Cette section détaille donc l'ensemble des entités qui constitue la HTMKS. Une distinc-

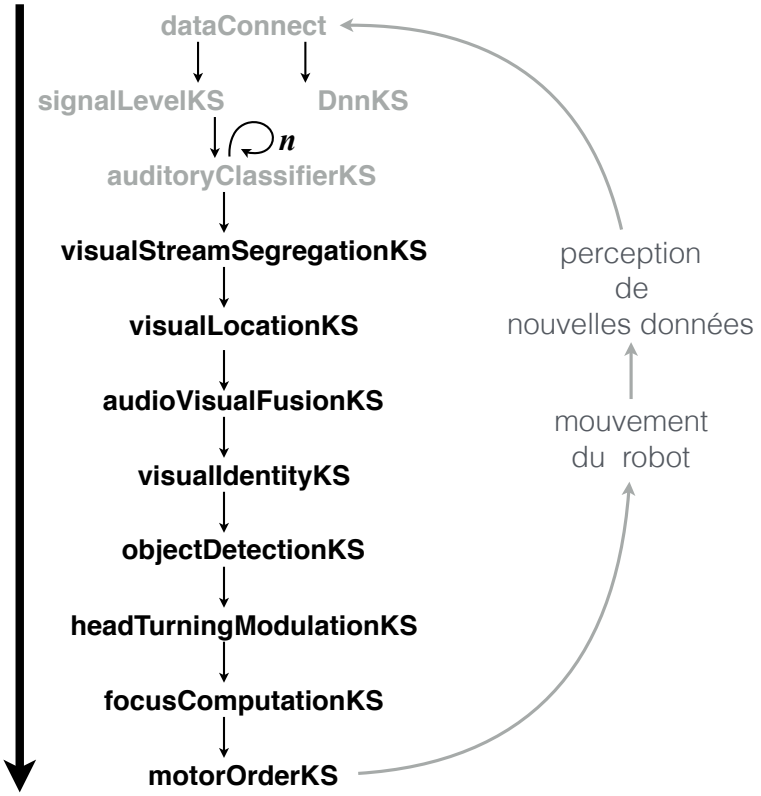


FIGURE 7.9 – CONNEXION DES DIFFÉRENTES KS — (*gris*) entités externes à la HTMKS, (*noir*) entités appartenant à la HTMKS. La flèche noire indique la connexion des différentes entités assurée par le SCHEDULER. La HTMKS, dans son ensemble, arrive en toute fin de chaîne, ayant un rôle plus cognitif et se basant sur les analyses de nombreuses autres KS. (*dataConnect*) structure permettant principalement la connexion entre le robot et TWO!EARS.

---

tion sera notamment faite (comme pour les autres entités implémentées au sein de TWO!EARS) entre les *Sources* et les *Knowledge Sources* : une Source est une entité conceptuelle n'effectuant pas d'analyse comme les experts le font et constituant plutôt la représentation d'une entité réelle, un objet audiovisuel ou le robot, par exemple. Dans ce qui suit, toutes les entités portant un nom terminant par « KS » signifie donc qu'elles ont été implémentées en tant que KS — les autres étant ainsi des Sources.

HEADTURNINGMODULATIONKS : premier point de communication avec le *Blackboard*, cette KS permet de déclencer les différents composants nécessaires au fonctionnement du modèle HTM global. Notamment, elle fait appel à la OBJECTDETECTIONKS à partir de laquelle elle va appeler la Source ROBOTINTERNALREPRESENTATION grâce à laquelle les données reçues du *Blackboard* vont pouvoir être intégrées et analysées jusqu'à aboutir à la requête d'un ordre moteur par la FOCUSCOMPUTATIONKS et la MOTORORDERKS.

ROBOTINTERNALREPRESENTATION (Source) : implémentation du ROBOT (cf. **Déf. 1**) incluant la représentation interne des environnements qu'il a exploré (objets audiovisuels détectés). Cette classe rassemble ainsi toutes les

données que le robot $\mathcal{R}$ a observées et analysées dans chaque environnement $e^{(i)}$ exploré. Cette classe regroupe également tous les résultats des analyses faites par le module DW.

PerceivedEnvironment (Source) : implémentation d'un Environnement (cf. **Déf. 5**). Cette classe contient tous les objets o_j détectés au sein de chaque environnement $e^{(i)}$. De plus, cette source contient une instance spécifique du module DW qui sera utilisée pour le mécanisme de transmission des connaissances (comme décrit à la **Sec. 5.3.3**). De plus, cette source est connectée directement avec le module MFI et est la principale voie de communication des données vers lui.

PerceivedObjet (Source) : implémentation d'un Objet (cf. **Déf. 4**), cette classe est la conjonction de toutes les données perçues par le robot pour un objet donné, grâce à sa connexion à la classe PerceivedEnvironment. Elle contient, entre autres, les angles $\theta_{a/v}$ estimés par les différents experts de localisation, les labels audio et visuels $(a/v)_j$ estimés par le module MFI et l'analyse de la Congruence estimée par le module DW. Un système de détection de données manquantes est également intégré à cette Source, permettant d'envoyer des requêtes entrant dans l'analyse de l'éventuelle nécessité d'un mouvement de tête vers l'Objet qu'elle représente. Enfin, elle contient la valeur de la pondération effectuée par le module DW.

AudioVisualKS : cette KS analyse superficiellement les Evénements (cf. **Déf. 3**) perçus par le robot, en particulier les informations de localisation audio et visuelles. Il s'agit de la première étape de la détection d'objet. Si plusieurs événements visuels sont détectés par le système visuel, plusieurs angles de localisation θ_v seront disponibles. Cette Knowledge Source a pour but d'assigner chaque localisation visuelle à la localisation audio correspondante, si elle existe, afin qu'il n'y ait pas de concaténation incorrecte des données lors de la création et la mise à jour des Objets.

VisualIdentity[QR]KS : cette KS est en charge de la récupération des données de reconnaissance des objets visuels détectés (images apprise pour le robot Jido et codes QR lus en temps réel pour le robot ODI). Cette KS permet de créer le vecteur $\mathbf{P}_v[t]$.

VisualLocationKS : cette KS s'occupe de la récupération des données de localisation visuelle lorsqu'un objet visuel a été détecté. Cette KS permet de créer le vecteur $\mathbf{\Theta}_v[t]$.

ObjectDetectionKS : sur la base des informations fournies par l'AudioVisualKS, cette KS va chercher si un objet situé dans la même zone de l'environnement a déjà été perçu ou, au contraire, s'il est nécessaire de créer un nouvel Objet *via* une nouvelle instanciation de la classe PerceivedObject.

FocusComputationKS : cette Knowledge Source rassemble les requêtes du module MFI et du module DW pour la génération de mouvements de

tête vers chaque OBJET présent au temps t dans un ENVIRONNEMENT. C'est au sein de cette KS que les activités des deux modules sont calculées (cf. **Sec. 7.1.1** pour l'activité modifiée du module DW et voir la **Sec. 6.4** pour l'activité du module MFI). Cette KS implémente également la notion de persistance temporelle.

MOTORORDERKS : cette KS va recouper la requête de la FOCUSCOMPUTATIONKS avec la position de la tête du ROBOT afin d'envoyer une requête au BLACKBOARD. Cette KS assure ainsi que l'ordre moteur est cohérent avec le repère du ROBOT et relativement à θ_0, l'angle de la position initiale du robot, utilisé lors des expérimentations.

DYNAMICWEIGHTING (Source) : implémentation du module DW, responsable de l'analyse de la Congruence des catégories audiovisuelles auxquelles les objets détectés dans un environnement donné appartiennent (cf. **Chap. 5**). Cette Source est responsable du calcul de la Congruence et de la pondération des objets. Elle est également directement connectée au module MFI afin d'avoir accès aux catégories audiovisuelles créées par lui. Il y aura une instance du DYNAMICWEIGHTING par PERCEIVEDENVIRONMENT afin de pouvoir effectuer la transmission de connaissances entre deux environnements.

MULTIMODALFUSIONANDINFERENCE (Source) : implémentation du module MFI, responsable de l'apprentissage de la scène audiovisuelle et de l'inférence de données manquantes, détaillée au **Chap. 6**. Cette Source est le lien entre l'instance du MULTIMODALSELFORGANIZINGMAP qu'elle intègre, et le reste de la HTMKS. Notamment, elle analyse les catégories audiovisuelles perçues du point de vue de sa connaissance suffisante ou non de leur apprentissage. Cette analyse sera effectuée conjointement avec le PERCEIVEDENVIRONMENT.

MULTIMODALSELFORGANIZINGMAP (Source) : implémentation du M-SOM, réalisant l'apprentissage des données perçues par le robot après un traitement superficiel effectué par le MULTIMODALFUSIONANDINFERENCE et le PERCEIVEDENVIRONMENT. Cette Source ne communique qu'avec le module MFI.

7.2.1.1 Discussion

La HTMKS est une large structure contenant tous les composants nécessaires pour faire fonctionner le modèle HTM sur un vrai robot, avec de vraies données audio et visuelles et s'appuyant sur l'architecture TWO!EARS. Cette implémentation a été une adaptation de la version utilisée en simulation à une version utilisable sur un vrai robot et avec de vraies données. Notamment, chaque KS envoie le résultat de ces analyses au *Blackboard* et à la HTMKS, agissant ainsi comme une seconde structure de type *Blackboard*. A l'issue de l'analyse de tous les composants du HTMKS, la donnée principale qui sera utilisée par le système TWO!EARS sera l'ordre moteur généré par la FOCUSCOMPUTATIONKS et la MOTORORDERKS.

La section suivante décrit maintenant succintement les particularités du robot Odi sur lequel le système Two!Ears contenant la HTMKS a été porté.

7.2.2 Intégration sur le robot Odi

Tous les résultats présentés jusqu'à présent ont été obtenus *via* le HtmTestBed. Cependant, tout au long de cette thèse, un travail conséquent a été fourni sur les véritables robots et notamment sur Odi. Nous avons d'ailleurs pris soin, lors du développement du modèle HTM, de toujours prendre en compte toutes les caractéristiques du robot ainsi que du logiciel Two!Ears intégré afin de pouvoir utiliser directement le modèle développé en simulation, sur le vrai robot.

Odi et Jido sont quasiment similaires. Cependant, quelques différences importantes sont à noter. Tout d'abord, Odi ne dispose pas de vision binoculaire, celle-ci ayant été développée spécifiquement pour Jido. A la place, nous avons utilisé une *webcam* disposant d'une résolution moindre et non dotée de reconnaissance d'objets similaire à celle dont Jido a été doté. Nous avons ainsi utilisé une reconnaissance visuelle basée sur des codes QR. La localisation, quant à elle, a été effectuée par combinaison des résultats de la détection de codes QR et des données odométriques du robot, notamment la position de la tête.

D'autre part, la vision binoculaire de Jido consiste en une paire de lunettes construite par impression 3D prenant en compte précisément la morphologie de la tête KEMAR. Pour Odi, nous avons dû placer la *webcam* sur le haut de sa tête. Cette position sur la tête a d'ailleurs un effet sur la perception visuelle : lorsque la tête effectue une rotation, un point au sommet de la tête d'Odi n'effectuera pas le même mouvement qu'un point situé au niveau des yeux. Durant la création des environnements de tests en conditions réelles, un effort conséquent a dû être fait sur la position précise des codes QR, sans quoi la webcam ne pouvait pas les reconnaître. De plus, les conditions de lumière (luminosité mais surtout type de lumière) ont eu un impact majeur et des lumières spéciales ont dues être utilisées.

Concernant les mouvements d'Odi, la base mobile utilisée a causé des difficultés pour faire se mouvoir le robot. Deux raisons à cela. La première est que cette base mobile, différente de celle de Jido, dispose de seulement trois roues (quatre pour Jido) entraînant un basculement du centre de gravité vers l'avant lorsque le torse et la tête KEMAR ont été fixés dessus. Ainsi, bien que capable de bouger dans l'environnement, ce mouvement peut entraîner une chute si un obstacle même petit est sur son chemin. Deuxièmement, le revêtement utilisé dans la pièce où les expériences ont été conduites gêne légèrement la rotation des roues entraînant (i) des erreurs dans la navigation dûes à une différence entre l'ordre moteur demandé et l'action motrice réellement effectuée, et (ii) un ralentissement du déplacement. Ces deux raisons ont entraîné notre choix d'effectuer des expériences sans utiliser le mouvement d'Odi et en plaçant donc toutes les sources dans un champ de 180° face à sa tête. Nous verrons, à la section présentant l'évaluation sur le robot, que ces limitations n'ont malgré tout pas empêché une évaluation pertinente du modèle.

Du point de vue du système auditif, celui-ci est le même que Jido. En revanche, l'ensemble des experts du système Two!Ears, en particulier ceux dédiés à la localisation et l'identification audio, ont été entraînés dans la salle d'expérimentations

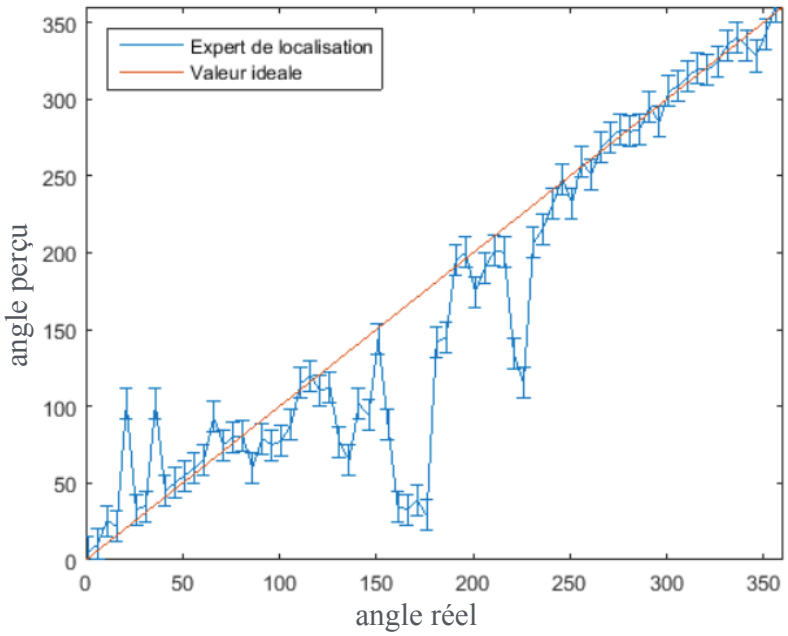


FIGURE 7.10 – TESTS DE L'EXPERT DE LOCALISATION — Effectué dans la salle de test de l'ISIR, ces résultats sont les moyennes de dix estimations successives de la position en azimut d'une source sonore se déplaçant par pas de 5°, effectuée par l'AUDIOLOCATIONKS utilisée en mai 2016. Une KS utilisant un apprentissage basé sur des réseaux de neurones profonds, la DnnLocationKS, est désormais utilisée. Ses performances sont légèrement meilleures.

---

du LAAS à Toulouse. Les conditions acoustiques sont extrêment différentes : la salle du LAAS est la réplication d'un appartement réaliste mais ne disposant pas de plafond (la hauteur au toit est d'ailleurs de près de 7 mètres) ; la salle est bien plus grande (plusieurs centaines de mètres carrés, contre une douzaine à l'ISIR). Etant données toutes ces différences, le comportement des experts audio sera susceptible d'être différent. Notre pièce étant quasi anéchoïque, cela devrait atténuer l'impact de cette différence. Nous avions d'ailleurs effectué quelques tests de localisation en mai 2016, illustrés à la **Fig. 7.10**, montrant plusieurs zones d'incertitude de localisation (hormis la confusion avant-arrière classique et pour laquelle les mouvements de tête participeront à sa résolution). Certaines améliorations ont été apportées depuis lors avec notamment l'utilisation de la DnnLocationKS (décrite à la **Sec. 3.2.3**), basée sur un apprentissage par réseaux de neurones profonds, mais nous n'avons pas encore pu effectuer de nouveaux tests. Cependant, durant les expériences menées dans le cadre de la validation du modèle, nous avons constaté une meilleure performance de la DnnLocationKS, bien que sujette à erreurs régulières, même sous la contrainte du placement des sources dans un champ de 180°.

7.2.3 Résultats sur Odi

Cette section présente les expériences réalisées sur le robot Odi et permettant d'observer le comportement global de la HTMKS en conditions réelles, utilisant des sons réels et des objets visuels sous formes de codes QR. Etant donné le temps

nécessaire au logiciel TWO!EARS (HTMKS incluse) pour effectuer l'analyse d'une trame (aux alentours de 2 s), chaque source émettra un son durant 15 s et une pause d'1 s sera systématiquement introduite avant une nouvelle émission. D'autre part, les experts d'identification et de localisation ayant été entraînés dans des conditions spécifiques (salle de test robotique du LAAS, à Toulouse), ceux-ci sont nettement moins performants dans notre salle de test de l'ISIR, malgré des conditions quasi anéchoiques. Ainsi, nous nous sommes pour le moment limité au cas unisource afin de minimiser l'impact du changement de lieu sur les différents experts. De plus, étant donné la configuration de la pièce dans laquelle nous avons effectué les expériences, toutes les sources ont dues être placées de façon similaire aux environnements simulés pour l'évaluation du module DW : dans un champ de 180° face au robot. Cependant, et contrairement à l'environnement simulé pour le module DW, le robot a un champ de vision d'environ 30° seulement. Nous avons de plus pris soin de disposer toutes les sources audiovisuelles en dehors de son champ de vision à l'état de repos, c'est-à-dire à minimum 30° à gauche et à droite, le forçant ainsi à effectuer un mouvement de tête pour accéder à la modalité visuelle de n'importe quelle source.

La première partie de cette section est dédiée à l'analyse du comportement global de la HTMKS en conditions réelles et dans un scénario à plusieurs environnements successifs tandis que la seconde partie est dédiée à l'analyse plus spécifique de la qualité de la classification effectuée par le module MFI.

7.2.3.1 Comportement global

Cette expérience a consisté en un scénario composé de trois environnements successifs différents, listés au **Tab. 7.4**. Ce scénario va nous permettre d'évaluer toutes les caractéristiques du modèle HTM : fusion de classifieurs, corrections d'erreurs, inférence de données manquantes et transmission de connaissances.

Caractéristiques du scénario de test 7.2.3.1 [6]					
$e^{(i)}$	$n_{\mathcal{S}}$	n_{sim}^{max}	Catégories présentes	$\theta^{(a\|v)}$	K_q
1	3	1	*dog barking* n°1 *dog barking* n°2 *female speech*	320° 35° 70°	0.6
2	3	1	*baby crying* n°1 *baby crying* n°2 *female piano*	70° 35° 320°	0.6
3	3	1	*baby crying* n°1 *baby crying* n°2 *dog barking* *male speech*	70° 35° 320° 280°	0.6

TABLE 7.4 – Caractéristiques du scénario en conditions réélles créé pour évaluer la HTMKS sur le vrai robot. Trois environnements successifs ont été créés permettant d'observer également le mécanisme de transmission ds connaissances.

6. Se référer à la **Sec. 4.3** pour l'explication de ces notations.

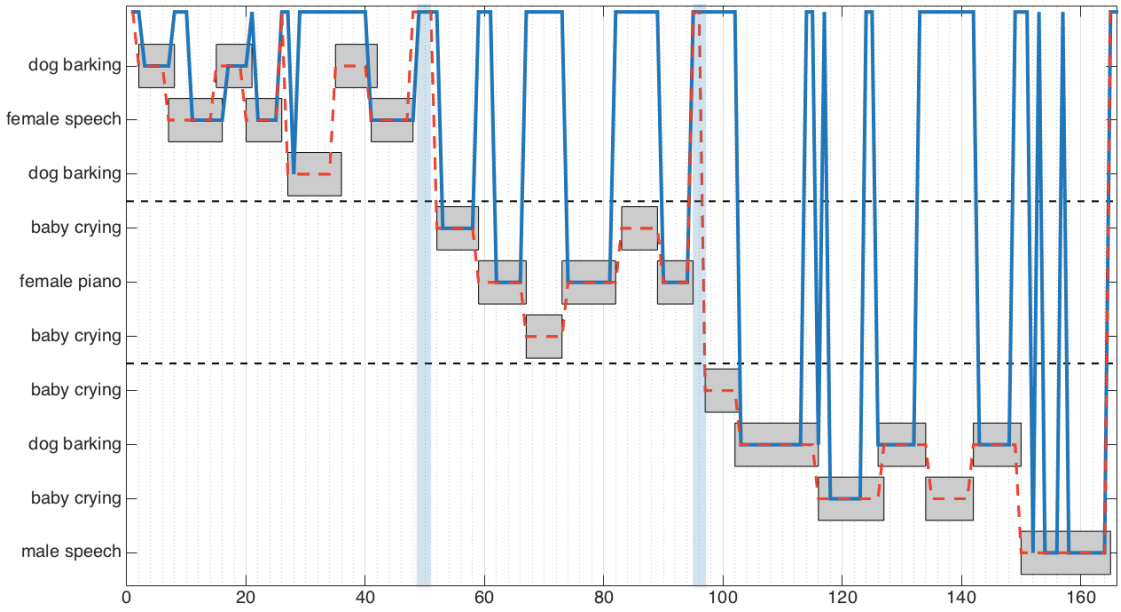


FIGURE 7.11 – OBJETS FOCALISÉS PAR LA HTMKS — (*ligne pleine bleue*) mouvements générés par la HTMKS, que ce soit par le module MFI ou le module DW, (*ligne rouge pointillée*) comportement du robot naïf virtuel. (*rectangles gris*) sources audiovisuelles, (*lignes noires pointillées*) et (*barres semi-transparentes verticales*) délimitations des différents environnement.

La **Fig. 7.11** présente les sources audiovisuelles qui ont été focalisées par le robot soumis au HTMKS (en bleu) dans les trois environnements successifs différents. Premièrement, nous pouvons observer l'impact du module DW à $t = 35$: la source $\mathcal{S}_1$ (*dog barking*) est ignorée par le robot puisqu'étant détectée comme congrue du fait de l'apparition de la source $\mathcal{S}_3$ (*dog barking*) un peu avant. D'ailleurs, nous notons une erreur à $t = 28$ mais qui a été corrigée à $t = 29$: la HTMKS a ordonné un ordre moteur (*via* le module DW) vers $\mathcal{S}_3$ supposé congru. Nous observons le même phénomène à $t = 57$ et $t = 83$. La rapidité avec laquelle le module DW génère des mouvements de tête (étant donné la priorisation du module MFI, comme décrit à la **Sec. 7.1**) nous permet de valider la pertinence de la façon dont les deux modules ont été combinés. Le module MFI parvient à apprendre rapidement les données perçues et laisse ainsi le module DW diriger le robot, lui permettant donc d'adopter un comportement attentionnel très tôt dans l'exploration d'environnements inconnus.

D'autre part, l'environnement n°3 nous permet d'observer le phénomène de transmission des connaissances : la première apparition de la source de catégorie *baby crying* est directement ignorée par la HTMKS. En effet, le système considère qu'il se situe dans un environnement comparable à $e^{(2)}$ dans lequel cette catégorie avait été considérée comme congrue. Dès lors que la deuxième source apparaît, de catégorie *dog barking*, la règle d'inclusion, exprimée par la **Déf. 13**, aboutit à la création d'un nouvel environnement : en effet, aucun environnement jusqu'alors exploré ne contient les deux catégories *baby crying* et *dog barking*. Cette dernière source est alors considérée comme incongrue et le module DW génère un mouvement de tête vers elle. Nous notons d'ailleurs ici une erreur de la part du module DW : le mouvement de tête généré à $t = 118$ vers la source n°3 aurait dû être inhibé. Il le sera

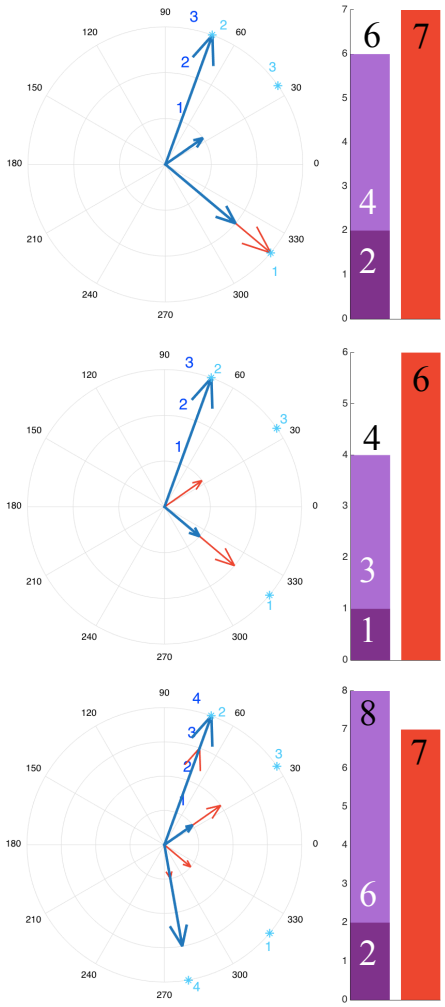


FIGURE 7.12 – MOUVEMENTS DE TÊTE EN CONDITIONS RÉELLES — llustration du nombre de mouvements de tête générés par (*bleu*) la HTMKS (*rouge*) le robot naïf virtuel. Chaque flèche pointe vers la position d'une source sonore et leur longueur dénote le nombre de mouvements vers la source pointée. (*histogrammes bleus*) nombre totaux de mouvements générés par (*violet foncé*) le module MFI, (*violet clair*) le module DW (les nombres en noir sont la somme des deux). (*histogrammes rouges*) nombre de mouvements générés par le robot naïf virtuel.

lorsque cette source se remettra à émettre, à $t = 132$.

Enfin, la **Fig. 7.12** présente, de façon similaire aux conditions simulées, le nombre de mouvements de tête générés vers chacune des sources présentes, par environnement. Ces nombres sont comparables au robot naïf virtuel tournant sa tête dès lors qu'un événement apparaît dans l'environnement, de façon identique au robot naïf utilisé pour les simulations. Nous voyons ici que ces nombres sont équivalents. Néanmoins nous observons un résultat important : le module DW prend très rapidement le pas sur le module MFI, notamment dans l'environnement n°3 où 8 mouvements de tête ont été requis par le module DW contre seulement 2 pour le module MFI. Les résultats pour cet environnement sont particulièrement intéressants, lorsque lus en parallèle de la **Fig. 7.11**. 5 mouvements ont été requis par le module DW pour les catégories *dog barking* et *baby crying*, de $t = 102$ à $t = 150$, deux catégories que le module MFI avait déjà rencontrées dans les deux environnements précédents. Ceci permet de voir que la HTMKS s'est servie des connaissances aquises précédemment pour analyser plus rapidement l'environnement actuel. D'autre part, lorsque la source *male speech* apparaît dans l'environnement, deux mouvements sont requis par le module MFI (dont un est causé par le retour à la position de repos à $t = 153$, mouvement « erroné ») afin d'apprendre cette nouvelle catégorie. Enfin, le dernier mouvement est requis par le module DW, la catégorie à laquelle ce dernier objet

appartient ayant été considérée comme incongrue à cet environnement.

Cette section nous a permis d'observer plusieurs points : comportement global de la HTMKS, nombre de mouvements de tête et transmission des connaissances. Les résultats obtenus, en conditions réelles désormais, nous permettent de voir que le modèle donne au robot la capacité d'apprendre et de réagir rapidement à l'apparition de sources audiovisuelles dans des environnements

7.2.3.2 Taux de bonne classification

Caractéristiques du scénario de test 7.2.3.2 [7]				
$n_{\mathcal{S}}$	n_{sim}^{max}	Catégories présentes ($\theta^{(a\|v)}$)	K_q	
5	1	*female speech* n°1	320°	0.6
		female speech n°2	30°	
		male piano	60°	
		dog barking	90°	
		baby screaming	280°	

TABLE 7.5 – Caractéristiques du scénario en conditions réelles créé pour évaluer la HTMKS sur le vrai robot permettant de mesurer le taux de bonne classification du modèle comparé à la fusion effectuée en sortie directe des classifieurs. L'émission d'une source dure 15 s avec une pause d'1 s avant l'émission de la prochaine source.

La première expérience est une validation de la capacité du modèle à effectuer une fusion correcte des données issues des experts d'identification. Un scénario comportant comportant $n_{\mathcal{S}} = 5$ sources sonores a été créé, en condition unisource, présenté au **Tab. 7.5**.

Bien que simple, comparé aux environnements utilisés en simulations, ce scénario permet déjà d'analyser la performance du modèle, et du module MFI en particulier, lorsqu'il est confronté à des données réelles. La **Fig. 7.13** illustre le taux moyen de bonne classification effectuée par le module MFI (en bleu) comparé à la fusion réalisée en sortie directe des classifieurs (en rouge [8]). Nous observons ici une amélioration significative : de $\bar{\mathbf{\Gamma}}_{\mathfrak{R}_n} = 37,9\%$ pour la fusion directe à $\bar{\mathbf{\Gamma}}^{a'}_{\text{MFI}} = 69,6\%$ pour le module MFI. Par rapport aux résultats obtenus en simulation, les conditions pour lesquelles nous retrouvons ces valeurs est :

- pour $\bar{\mathbf{\Gamma}}^{a'}_{\text{MFI}}$: condition n°12 du cas multisource (cf. **Tab. 6.3**) avec $n_{\mathcal{S}} = 5$ sources sonores et $n_{sim}^{max} = 5$ sources émettant simultanément. Le résultat était alors de $71,3\%$ de bonne classification ;
- pour $\bar{\mathbf{\Gamma}}_{\mathfrak{R}_n}$: condition n°5 du cas multisource également avec $n_{\mathcal{S}} = 3$ sources sonores et $n_{sim}^{max} = 3$ sources émettant simultanément. Le résultat était alors de $41,4\%$.

Cette comparaison nous permet de mettre en avant deux points. Le premier, auquel on pouvait s'attendre, est que le modèle HTM offre de bien meilleures performances dans l'environnement de simulation, même si nous avons porté un soin particulier

7. Se référer à la **Sec. 4.3** pour l'explication de ces notations.
8. nous rappelons qu'en cas unisource, $\bar{\mathbf{\Gamma}}_{\mathfrak{R}_n} = \bar{\mathbf{\Gamma}}'_{\mathfrak{R}_n}$.

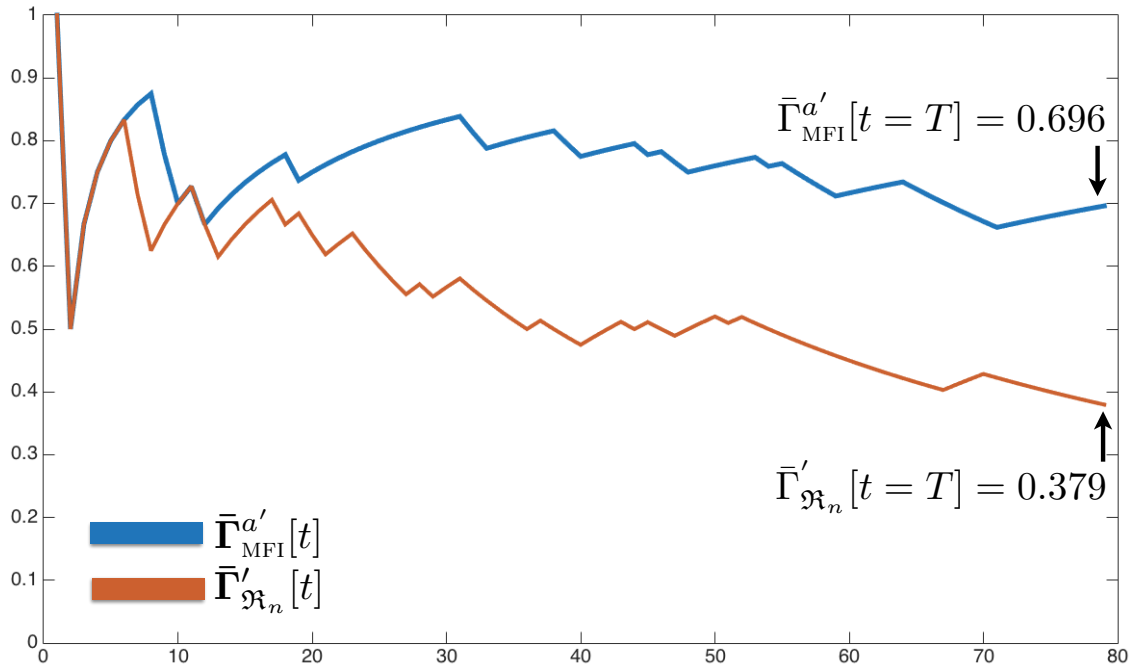


FIGURE 7.13 – TAUX DE BONNE CLASSIFICATION — Résultats de la classification audiovisuelle (incluant l'inférence pour le module MFI) obtenus par (*bleu*) la HTMKS, (*rouge*) le robot naïf virtuel. Les deux nombres correspondent à la valeur finale.

à rendre les conditions de simulation difficiles, que ce soit par le nombre de sources simultanées que par la façon dont les données sont simulées notamment avec le taux d'erreur $\varepsilon_{\surd}$. Cependant, il est important de mentionner que nous sommes ici en cas unisource. Or nous avons vu l'intérêt des capacités d'inférence du module MFI, et sa qualité. L'inférence est tout de même présente, notamment lorsque le module inhibe la génération de mouvements de tête une fois qu'il fait confiance en sa connaissance d'une catégorie, mais en proportion moindre que dans les conditions simulées. Ceci s'explique par des expériences plus courtes, en nombre de trames mais pas en temps « réel », du fait de la longueur du traitement des données par le système TWO!EARS entier. Le second point est que le modèle HTM permet quand même une nette amélioration de la qualité des données perçues par le robot et analysées par les différents experts et ce, dans des temps relativement courts : la **Fig. 7.13** nous montre que la qualité de la classification (fusion & inférence) n'est pas dépendante du temps. En effet, dès le début de l'expérience, le modèle parvient très rapidement à catégoriser correctement les données reçues. Nous mettons d'ailleurs ici en avant le fait que certaines catégories sont beaucoup mieux reconnues par les experts d'identification que d'autres, notamment la catégorie audio *femaleSpeech*. A l'inverse, certaines catégories sont plus sujettes à erreur : lors de l'émission d'un son de classe *barking*, l'expert dédié à la reconnaissance de la classe *alarm* montre souvent des probabilités d'appartenance très élevées, prenant même le pas sur toutes les autres.

Mais l'action du module MFI a un impact beaucoup plus grand sur les données perçues. La **Fig. 7.14** illustre toutes les catégories audiovisuelles différentes qui ont été le résultat des deux types de fusion. Nous voyons là que le module MFI ne se

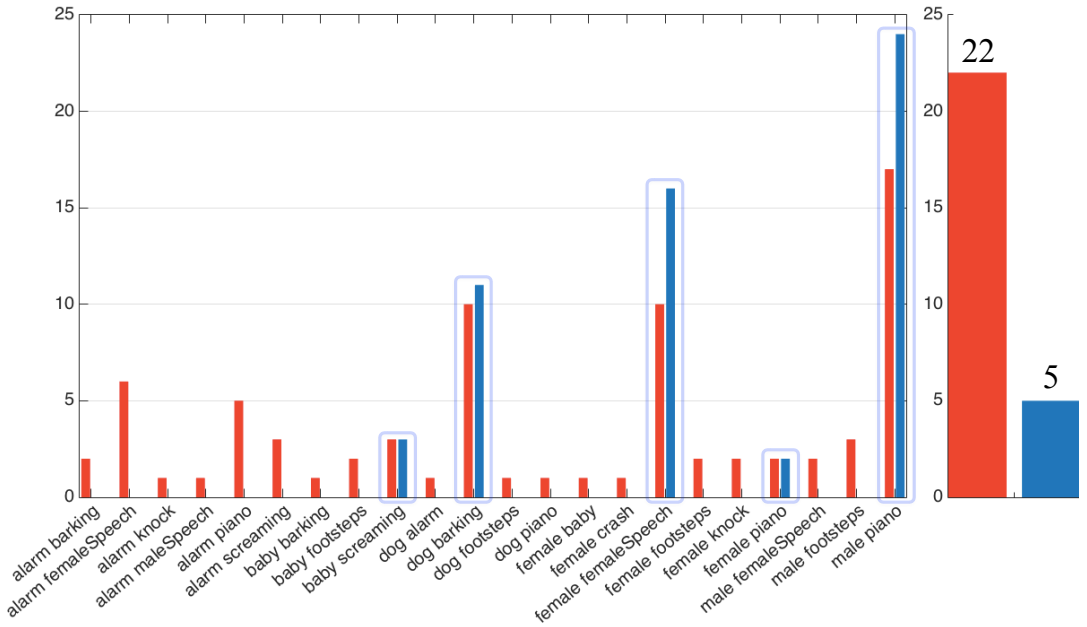


FIGURE 7.14 – CATÉGORIES AUDIOVISUELLES CRÉÉES — Illustration de la différence entre le système de fusion « naïf » et le module MFI du point de vue du nombre de catégories audiovisuelles différentes créées par (*rouge*) le système naïf, (*bleu*) le module MFI. (*gauche*) La hauteur des histogrammes représente le nombre de trames pour lesquelles la catégorie audiovisuelle considérée a été le résultat de la fusion des experts. Les rectangles bleu clairs mettent en avant les catégories audiovisuelles communes entre les deux systèmes. (*droite*) nombre total de catégories audiovisuelles différentes détectées.

contente pas de corriger d'éventuelles erreurs des experts d'identification : tirant parti de l'expérience passée du robot, il parvient à réduire grandement l'espace des possibles sur lequel le module DW va se baser. En effet, là où la fusion directe propose jusqu'à 22 catégories audiovisuelles différentes, le module MFI lui n'en a détecté que 4 différentes. De plus, une des catégories que le module MFI a détectée est *female piano*, combinaison de *female speech* et *male piano*. Ce mauvais appariement provient de la combinaison des informations audio perçues d'une source située à une position des informations visuelles perçues d'une autre source devant laquelle le robot était situé.

D'autre part, la dynamique temporelle du robot et des traitements effectués par le logiciel TWO!EARS (HTMKS incluse) compliquent la tâche de la HTMKS. Nous mettons en avant deux phénomènes gênant l'analyse temps-réel des stimuli perçus par le robot et aboutissant ainsi à des situations difficiles à gérer pour notre modèle :

1. Certaines trames prennent jusqu'à 2 s pour être traitées. Durant ce temps de traitement, la scène est susceptible de grandement changer. Allonger les temps d'émission de chaque source n'est pas une solution puisque le problème vient du moment où il y a un changement d'activité des sources audiovisuelles ;
2. Si justement un changement d'activité des sources survient dans l'intervalle de capture des stimuli (une trame temporelle dure 500 ms), le résultat de l'analyse des experts intégrera des informations sur toutes les sources ayant été perçues à ce moment-là.

Malgré cela, la gestion des données faite par la HTMKS, notamment tentant de rassembler les événements perçus au sein de la notion d'Objets, fait que notre système parvient à une analyse pertinente d'un environnement réel.

7.2.3.3 Discussion

Cette section a été dédiée aux premiers résultats obtenus sur la plateforme robotique de l'ISIR, Odi. Nous avons créé différents scénarios permettant d'évaluer le comportement et les performances de la HTMKS, version du modèle HTM intégrée à l'ensemble du système Two!Ears. Ce passage au monde réel constitue la véritable validation du modèle en cela que ce monde, avec de vrais signaux audiovisuels analysés en ligne et un vrai robot réagissant en temps réel, est beaucoup plus difficile à comprendre que le monde simulé. Néanmoins, malgré l'instabilité des données issues des experts d'identification (et de l'expert de localisation audio, même si nous n'avons pu faire d'évaluation de sa performance dans notre pièce), la HTMKS a montré sa capacité à conférer au robot :

1. un système d'apprentissage très rapide des données réelles permettant une fusion et une inférence des données performantes, apprentissage effectué sans connaissance *a priori* ;
2. une compréhension de l'environnement sous forme d'Objets ;
3. un comportement attentionnel basé sur la Congruence.

En revanche, les conditions dans lesquelles nous nous sommes placés pour effectuer cette validation seront à complexifier : cas multisource d'une part et, d'autre part, dès que la plateforme robotique le pourra, position des sources à 360°. Enfin, faire évoluer le robot dans un véritable environnement complexe dans lequel il analyse la scène audiovisuelle grâce au modèle HTM sera une étape indispensable afin de valider complètement le modèle.

7.3 Conclusion du Chapitre

Ce chapitre a concerné la description de l'état final du modèle HTM. Premièrement, la combinaison des deux modules a été présentée et l'évaluation de cette version complète du modèle a été effectuée en simulation. Nous avons notamment vu la dynamique des activités du module MFI et du module DW, le second n'étant actif que lorsque le premier devient inactif. Cette collaboration entre les deux modules permet au modèle de donner au robot un comportement pertinent : les données perçues par le robot sont d'abord traitées, les éventuelles erreurs sont corrigées et la fusion des classifieurs est effectuée, puis ces données sont envoyées au module DW qui, quant à lui, nécessite des données robustes afin de prendre une décision pertinente sur la congruence ou l'incongruence des objets auxquels ces données sont associées.

D'autre part, nous avons décrit une partie de l'adaptation du modèle HTM en son équivalent Two!Ears : la HTMKS. Celle-ci peut être vue comme un sous-*Blackboard* agissant comme l'organisateur de toutes les entités composant cette KS. Notamment, nous avons vu que les définitions faites à la **Sec. 4.1** ont toutes trouvées leur

équivalent computationnel. L'adaptation du modèle en KS a permis de l'utiliser en conditions réelles, c'est-à-dire sur Odi, la plateforme robotique de l'ISIR, et de pouvoir récupérer les vraies données issues des experts TWO!EARS. Après avoir détaillé les contraintes particulières de notre robot, en comparaison de Jido, le robot du LAAS, nous avons présenté les résultats obtenus dans divers conditions de test. Ces résultats sont très encourageants puisque nous retrouvons ceux obtenus en simulation. L'intégration sur un robot réel est en effet une étape cruciale : l'environnement de simulation HtmTestBed, aussi contraignant soit il dans sa façon de simuler des environnements, ne s'approche que peu des conditions réelles. Cette intégration a malgré tout été un succès, même si nous avons dû prendre en compte, lors de la création des environnements de test en conditions réelles les limitations actuelles de notre robot. Nous continuons ainsi de travailler sur la validation du modèle en conditions de plus en plus complexes et réalistes.

Arrivé à ce stade néanmoins, nous pouvons dire que le modèle HTM entier a été validé en conditions simulées et en conditions réelles.

Chapitre 8

Conclusion

Ce chapitre final clôt ce manuscrit, concrétisation de la majeure partie du travail effectué durant cette thèse. Il consistera d'abord en la description des limites de chaque module du modèle ainsi qu'en les pistes de recherche future que nous avons déjà commencé à explorer. Nous terminerons par une conclusion globale sur tout ce qui a été présenté dans ce manuscrit.

8.1 Limites du DW et travail futur

Une des principales limites du module DW est que nous n'avons pas pris en compte l'aspect temporel dans notre définition de la Congruence. En effet, nous avons considéré le module DW comme étant un système focalisé sur la notion d'événement, au sens d'apparition d'un stimulus dans une zone de l'espace, plutôt que sur celle de durée d'un stimulus. Il n'y a donc pour le moment aucun phénomène d'habituation intégré à la fonction de pondération. Cependant, cette habituation existe et a été observée et mesurée chez l'humain : il s'agit même d'une forme de saillance. Un stimulus d'intérêt peut occasionner une réaction motrice du fait de son incongruence mais passera progressivement en arrière-plan lorsque le contenu informationnel sera de plus en plus faible. Nous proposerons à la **Sec. 8.3** une piste prometteuse pour inclure cette caractéristique de la Congruence d'un événement.

D'autre part, bien que permettant déjà une analyse puissante d'environnements inconnus, le module DW à lui seul n'est que très imparfait pour rendre compte d'une réaction attentionnelle riche et pertinente. Cela n'a d'ailleurs jamais été son ambition : ce module est une source de connaissance supplémentaire devant être utilisée conjointement avec d'autres sources de connaissances. Malheureusement, au sein du projet Two!Ears, aucune KS similaire n'a pu être développée et intégrée au robot, le modèle HTM et son composant attentionnel, le module DW, étant ainsi les briques les plus « haut-niveau » du projet. Or il serait intéressant d'inclure d'autres caractéristiques dans la définition de ce qu'est un Objet, afin de l'enrichir, de rendre chaque objet plus spécifique et plus précis, dans le but d'affiner la réaction comportementale initiée par le module DW. En effet, tout part de la définition d'un Objet

et cette définition peut être largement complétée : caractéristiques audio plus poussées (timbre, hauteur, émotion...) ou visuelles (couleur, forme, texture...), mais également à partir d'autres modalités comme des données tactiles par exemple. A partir de cette définition, le calcul de la Congruence serait plus fin, plus pertinent. Une des propositions d'extension du M-SOM inclut d'ailleurs l'intégration de données spatiales dans des vecteurs de poids distincts, permettant ainsi de définir un objet en utilisant une nouvelle source d'information.

D'autre part, l'ensemble du module DW repose sur de bons résultats de localisation fournis notamment par les experts audio. En effet, sans cette information, aucun mouvement de tête ne peut être généré correctement. Nous avons vu que même avec le robot réel nous avons pu obtenir un comportement similaire à celui observé en simulation, malgré des résultats de localisation parfois erronés. En addition, les experts de localisation, tout comme ceux d'identification, ont été entraînés dans des conditions très précises rendant difficiles l'utilisation du robot dans des environnements différents de ceux qui ont été appris. Ainsi, il serait extrêmement bénéfique pour le module DW de pouvoir également avoir accès à des données spatiales plus précises, que ce soit par une étape d'analyse supplémentaire après celle des experts, ou par un raffinement de l'apprentissage effectué par les experts.

8.2 Limites du MFI et travail futur

Le module *Multimodal Fusion & Inference* permet d'apprendre le lien entre les modalités utilisées pour définir ce qu'est un Objet. Dans notre cas, nous avons utilisé les informations audio et visuelles issues des experts d'identification fournis par le système Two!Ears. Nous avons étudié les performances du module MFI à la **Sec. 6.5**, notamment sa capacité à inférer de l'information à partir d'une information incomplète ainsi que sa gestion des erreurs de classification des experts d'identification.

Une des principales limites, de façon similaire au module DW, est qu'il repose intensément sur la position spatiale perçue des sources audiovisuelles peuplant l'environnement en cours d'exploration. Sans cette information, le module MFI ne peut pas non plus fonctionner correctement puisque les mouvements de tête requis n'auront aucune information sur la position de la source cible.

Nous nous sommes alors penchés sur la possibilité d'intégrer les données de localisation audio et visuelle, obtenus par les experts dédiés, dans le M-SOM, créant ainsi deux nouveaux vecteurs de poids et permettant de définir un objet également par sa position spatiale. Ainsi, il serait possible de corriger les éventuelles erreurs de localisation ainsi que d'ajouter une nouvelle source d'information possiblement utilisable pour effectuer un nouveau type d'inférence de données manquantes. Les résultats préliminaires que nous avons obtenus en simulation sont extrêmement encourageants : il semble possible de gérer conjointement des taux d'erreurs conséquents touchant l'identification et la localisation audio et visuelle, grâce à l'utilisation du M-SOM. Nous avons alors pu formaliser la notion d'Objet en tant qu'entité audiovisuelle et spatiale directement au sein du M-SOM. Cependant, cette piste nécessite de poser une hypothèse forte, que nous avons d'ailleurs déjà dû poser durant cette

thèse : les sources sont statiques. Nous pensons cependant pouvoir passer outre cette contrainte en introduisant un suivi spatial des sources audiovisuelles au sein du M-SOM.

Enfin, une autre limite est celle de l'appariement un à un des classes audio et visuelles. Nous avons déjà justifié ce point précédemment mais nous pensons qu'il est possible — et même nécessaire — de pouvoir apporter une solution à ce problème. Plutôt que de prendre une décision sur l'appartenance à une catégorie audiovisuelle, il serait peut-être plus pertinent d'émettre une hypothèse sur cette appartenance. Or cette hypothèse peut également se retrouver dans le M-SOM directement : par l'analyse des distances calculées entre les vecteurs de poids et un vecteur d'entrée à catégoriser. Plaçons nous dans un cas d'inférence où seule la donnée audio est disponible et où cette classe audio correspond à deux classes visuelles différentes. Cela signifie qu'il existera certainement deux zones du sous-réseau audio pour lesquelles les distances seront particulièrement faibles. Par analyse des classes visuelles correspondantes à ces deux zones (puisque nous avons effectué un apprentissage sous une contrainte de dépendance des deux sous-réseaux), il serait possible d'émettre une hypothèse sur les deux catégories audiovisuelles auxquelles le vecteur considéré est susceptible d'appartenir. La décision pourrait alors être prise par le module DW : si une de ces deux catégories audiovisuelles est considérée comme incongrue, un mouvement de tête sera généré. Même si la catégorie réelle se trouve différente, l'acquisition de nouvelles données complètes permettra de relancer une itération d'apprentissage et ainsi peut-être de mieux capturer la différence qui existe entre ces deux catégories audiovisuelles à partir d'une donnée audio (peut-être la distribution entière des probabilités des experts d'identification, bien que celles-ci étant indépendantes entre elles, porte une différence en fonction de la catégorie visuelle correspondante). Cela conduirait également à l'élaboration d'un système de méta-information entrant dans la définition d'une catégorie : telle catégorie audio peut correspondre à telles catégories visuelles (et vice-versa). Ce système permettrait d'affiner l'inférence et de conférer au module MFI une analyse plus subtile et pertinente de données ambigües.

8.3 Unification du modèle

Le module DW est un module permettant de pondérer l'apparition d'un objet appartenant à une catégorie audiovisuelle par la notion de Congruence à l'environnement dans lequel il se situe. Cette Congruence est dépendante de la probabilité *a posteriori* de cette catégorie d'apparaître dans l'environnement considéré. Or nous disposons, au sein du modèle HTM, d'une structure dont le comportement possède déjà un comportement proche de la Congruence : le M-SOM.

En effet, plus une catégorie audiovisuelle sera perçue par le robot, plus elle sera apprise et plus la zone du M-SOM dédiée à la représentation de cette information grandira. Ce comportement, directement hérité du fonctionnement SOM, peut être utilisé pour modéliser la Congruence en cela que la détermination du nombre de neurones codant une catégorie comparé au nombre total de neurones peut être une formalisation de la Congruence. Ainsi, il serait possible d'intégrer le fonctionnement du module DW au sein du module MFI en une architecture unifiée permettant en

même temps d'apprendre l'environnement, du point de vue de son contenu sémantique, et de réagir en fonction de ce contenu.

Cependant, une limite apparaît ici : la Congruence est calculée par environnement et nous avons précisé que le M-SOM est une structure du robot, commune donc à tous les environnements explorés. Nous avons alors pensé à deux solutions : (i) à l'instar de la création d'un module DW par type d'environnement exploré, nous pourrions créer un SOM par environnement, communiquant directement avec le M-SOM et dont la propagation de l'apprentissage permettrait de modéliser la Congruence, ou (ii) d'intégrer les probabilités *a posteriori* d'apparition des catégories dans un vecteur de données qui correspondrait à un nouveau vecteur de poids du M-SOM et où chaque composante serait un environnement exploré.

8.4 A propos du robot naïf

Au sujet de cette comparaison, nous tenons ici à préciser un point important. Le robot naïf ne constitue qu'une référence nous permettant de juger de la pertinence des mouvements de tête effectués par notre modèle. Ce robot, pour rappel, est dirigé par une forme de motivation par la *Nouveauté*, motivation décrite à la **Sec. 2.1.3**, entraînant le robot à tourner sa tête à chaque fois qu'un événement audio ou visuel survient. Cette motivation peut également être une interprétation d'une forme de réaction à la saillance audio d'un événement : d'après Duangudom & Anderson [141] que nous citions à la **Sec. 2.3.1.2**, les sons *nouveaux* ont tendance à être considérés comme saillants par le système auditif tandis que les sons constants ou plus anciens tendent à passer dans à l'arrière-plan. A la lumière des mécanismes cérébraux que nous avons exposés sur la perception, l'integration multimodale et les réactions attentionnelles, nous savons que les aires sensorielles sont particulièrement sensibles à un stimulus saillant, qu'il le soit par ses caractéristiques bas-niveau ou son contenu sémantique. La saillance peut être aussi portée par des événements imprédictibles puisque la notion d'arrière-plan peut également être comprise comme *ce qui peut être prédit*, ce qui étant prédictible ne nécessitant ainsi plus toutes les capacités analytiques des caractéristiques bas-niveau des signaux perçus, comme la Théorie de la Hiérarchie Inverse le stipule. De plus, nous avons vu qu'une réaction particulière, la *Mismatch Negativity*, est observée lors de l'apparition de stimuli imprédictibles et qu'une réaction motrice peut s'en suivre.

Ainsi, bien que nous avons, tout au long de ce manuscrit, employé le terme de *naïf*, ce robot virtuel, que nous avons utilisé comme point de référence, ne l'est pas tant que ça, tout du moins du point de vue de sa réaction « comportementale ». En revanche, le fait qu'il ne dispose pas de mécanismes d'intégration temporelle des données qu'il perçoit ou de système d'apprentissage le rend effectivement naïf. Comme soulevé lors de l'évaluation en simulation du modèle HTM entier, la comparaison du nombre de mouvements de tête que nous avons effectuée à chaque fois qu'il était question de juger la pertinence du comportement du module DW et du module MFI n'est pas une preuve de succès. Nous avons en effet choisi, sur la base de la synthèse d'un ensemble de connaissances sur les mécanismes cérébraux impliqués, de diriger un robot exploratoire doté de mouvements de tête selon une détermination de la Congruence d'un événement apparaissant dans un environnement. Bien que puisant largement

son inspiration dans les phénomènes biologiques, il reste néanmoins certain que des comportements réellement pertinents, comme ceux observés chez l'animal, nécessitent des structures computationnelles bien plus larges, complexes et dotées de capacités que les robots actuels sont loin de posséder. Ainsi, utiliser ce robot naïf nous a malgré tout permis d'avoir une idée sur la pertinence du comportement du robot soumis au module DW, au module MFI, puis au deux ensemble.

8.5 Conclusion du Chapitre

Le modèle Head Turning Modulation présenté ici est la concrétisation de ces trois années de thèse. Développé dans le cadre du projet européen Two!Ears, il a permis de donner une étincelle de conscience (selon la définition d'Antonio Chella & Riccardo Manzotti [3]) à un robot mobile et doté de mouvements de tête, dans le cadre de l'exploration d'environnements inconnus. Nous avons pris soin d'exposer au mieux les concepts, théories et paradigmes qui ont inspirés la création de ce modèle. Ils sont nombreux : perception, intégration multimodale, exploration (localisation, navigation et cartographie), mécanismes attentionnels et apprentissage machine. Malgré tout, nous avons pu effectuer une synthèse de tous les champs de recherche impliqués lors de l'élaboration d'un système dont nous pouvons dire qu'il entre dans la catégorie d'« intelligence artificielle ». La combinaison de toutes ces connaissances sur le fonctionnement cérébral nous a permis de développer un modèle computationnel toujours ancré dans de solides considérations biologiques. Nous voyons d'ailleurs, après la validation en conditions simulées puis en conditions réelles, que l'implémentation de modules relativement simples permet déjà de faire émerger des comportements complexes et pertinents.

En introduction de ce manuscrit, nous avions posé la question suivante :

> *Comment faire en sorte qu'un robot mobile doté d'une perception audiovisuelle similaire à l'Homme puisse de lui-même comprendre ce qu'est un* stimulus d'intérêt *afin de porter son attention dessus, tout en restant sensible à tous les autres stimuli présents dans un environnement inconnu en cours d'exploration ?*

A cette question, nous avons répondu en utilisant premièrement la notion de *Congruence* d'un objet au sein de l'environnement dans lequel il est situé. Cette notion, inspirée de la saillance, bien définie dans le domaine audio mais qui a surtout été intensivement étudiée au sein de la communauté de la vision — qu'elle soit des neurosciences ou de la robotique — nous a permis de faire émerger un comportement attentionnel basé uniquement sur l'appartenance à une catégorie audiovisuelle et sans règles de comportement données *a priori*, comme c'est souvent le cas pour d'autres modèles attentionnels.

Les limites du module *Dynamic Weighting* responsable de ce calcul de la Congruence ont motivé le développement d'un second module : le module de *Multimodal Fusion & Inference*. Le problème alors posé était d'apprendre à inférer une information éventuellement manquante sur la base de données acquises en ligne et toujours avec

la contrainte d'absence de règles *a priori*. Ainsi, suivant les mêmes limitations que pour le module DW, nous sommes parvenus à élaborer un système d'analyse des données audiovisuelles d'un environnement inconnu permettant de façon robuste, rapide et non supervisée, de conférer au robot la capacité d'inférer une modalité à partir d'une autre (notamment l'information visuelle à partir de l'audio).

Que ce soit dans le cas du module DW ou dans celui du module MFI, les mouvements de tête ont été la pierre angulaire du modèle HTM. Ces mouvements de tête, dont les robots Jido et Odi sont dotés, multiplient les capacités d'analyse de l'environnement par ajout d'un degré de liberté au robot : celui-ci peut aller d'un point à un autre tout en acquérant de nouvelles données sur ce qui se passe autour de lui. Ne gênant ainsi pas les tâches de navigation, tâches courantes et essentielles en robotique mobile, ces mouvements de tête à eux-seuls sont un pas vers l'élaboration de robots humanoïdes tendant vers la compréhension du monde telle que nous, humains, l'avons. Cela a d'ailleurs été une des motivations du projet TWO!EARS qui s'est concrétisée par le choix de robots disposant d'audition binaurale, de vision binoculaire (du moins pour Jido) et d'un cou pouvant donner à la tête la possibilité de tourner. Le choix d'une audition binaurale entre d'ailleurs dans la même réflexion globale sur l'intérêt de créer des algorithmes d'intelligence artificielle, au sens large, intégrés à des robots. Car une des ambitions voilées de ces travaux de recherche est la même que celle des chercheurs en neurosciences : comprendre le fontionnement du cerveau.

Le modèle HTM se démarque de nombreux autres en cela qu'il se base sur des principes simples mais suffisamment ancrés dans un ensemble d'observations sur les mécanismes cérébraux pour atteindre une réelle efficacité dans sa compréhension de l'environnement. Un des fils rouges de cette thèse a été de considérer le problème à traiter non pas en essayant d'utiliser à tout prix les techniques et algorithmes les plus avancés de la recherche actuelle, mais en essayant de comprendre le véritable cœur du problème et de tenter d'y apporter, dans un premier temps, une solution assez simple pour pouvoir ensuite être enrichie incrémentalement. Car nous sommes convaincus que le modèle HTM pourrait avoir vocation à être largement étendu, et des pistes ont d'ailleurs déjà été proposées aux sections précédentes.

Bibliographie

[1] M. Corbetta, G. Patel, and G. L. Shulman, "Review The Reorienting System of the Human Brain : From Environment to Theory of Mind," pp. 306–324, 2008.

[2] A. Chella and R. Manzotti, "Machine Consciousness : a Manifesto for Robotics," *International Journal of Machine Consciousness*, vol. 01, no. 01, pp. 33–51, 2009.

[3] A. Chella and R. Manzotti, "Artificial Intelligence and Consciousness," in *Artificial Intelligence* (A. Chella and R. Manzotti, eds.), pp. 37–45, 2013.

[4] J. O'Keefe and L. Nadel, *The Hippocampus as a Cognitive Map*. No. 9, Oxford : Clarendon Press, 1978.

[5] L. Macedo and A. Cardoso, "Modeling Forms of Surprise in an Artificial Agent," in *Proceedings of the Cognitive Science Society*, vol. 23, 2001.

[6] M. A. Brown and P. E. Sharp, "Simulation of Spatial Learning in the Morris Water Maze by a Neural Network Model of the Hippocampal Formation and the Nucleus Accumbens," *Hippocampus*, vol. 5, no. 3, pp. 171–188, 1995.

[7] M. E. Ragozzino, S. Detrick, and R. P. Kesner, "Involvement of the Prelimbic-Infralimbic Areas of the Rodent Prefrontal Cortex in Behavioral Flexibility for Place and Response Learning," *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, vol. 19, pp. 4585–94, jun 1999.

[8] C. Lever, T. Wills, F. Cacucci, N. Burgess, and J. O'Keefe, "Long-Term Plasticity in Hippocampal Place-Cell Representation of Environmental Geometry," *Letters to Nature*, vol. 416, pp. 90–94, 2002.

[9] S. M. Nicola, I. a. Yun, K. T. Wakabayashi, and H. L. Fields, "Cue-Evoked Firing of Nucleus Accumbens Neurons Encodes Motivational Significance During a Discriminative Stimulus Task," *Journal of Neurophysiology*, vol. 91, pp. 1840–65, may 2004.

[10] V. Hok, E. Save, and B. Poucet, "Coding for Spatial Goals in the Prelimbic-Infralimbic," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 102, no. 12, pp. 4602–4607, 2005.

[11] S. a. Taha, S. M. Nicola, and H. L. Fields, "Cue-Evoked Encoding of Movement Planning and Execution in the Rat Nucleus Accumbens," *The Journal of physiology*, vol. 584, pp. 801–18, nov 2007.

[12] N. Cuperlier, M. Quoy, and P. Gaussier, "Neurobiologically Inspired Mobile Robot Navigation and Planning," *Frontiers in Neurorobotics*, vol. 1, 2007.

[13] J. O'Keefe and J. Dostrovsky, "The Hippocampus as a Spatial Map. Preliminary Evidence From Unit Activity in the Freely-Moving Rat," *Brain research*, vol. 34, no. 1, pp. 171–175, 1971.

[14] K. J. Jeffery and R. Hayman, "Plasticity of the Hippocampal Place Cell Representation," *Reviews in Neuroscience*, vol. 15, pp. 309–331, 2004.

[15] R. U. Muller and J. L. Kubie, "The Effects of Changes in the Environment Hippocampal Cells on the Spatial Firing of," *Journal of Neuroscience*, vol. 7, no. 7, pp. 1951–1968, 1987.

[16] G. J. Quirk, R. U. Muller, and J. L. Kubie, "The Firing of Hippocampal Place Cells in the Dark Depends on the Rat's Recent Experience," *Journal of Neuroscience*, vol. 10, no. 6, pp. 2008–2017, 2008.

[17] E. J. Markus, C. A. Barnes, B. L. McNaughton, V. L. Gladden, and W. E. Skaggs, "Spatial Information Content and Reliability of Hippocampal CA1 Neurons : Effects of Visual Input," *Hippocampus*, vol. 4, no. 4, pp. 410–421, 1994.

[18] J. L. Kubie and J. Ranck, "Sensory-Behavioral Correlates in Individual Hippocampus Neurons in Three Situations : Space and Context," in *Neurobiology of the Hippocampus*, pp. 433–447, 1983.

[19] R. E. Hampson, C. J. Heyser, and S. A. Deadwyler, "Hippocampal Cell Firing Correlates of Delayed-Match-to-Sample Performance in the Rat," *Behavioral Neuroscience*, vol. 107, no. 5, pp. 715–739, 1993.

[20] B. G. Young, G. D. Fox, and H. Eichenbaum, "Correlates of Hippocampal Complex-Spike Cell Activity in Rats Performing a Nonspatial Radial Maze Task.," *The Journal of Neuroscience*, vol. 14, no. 11, pp. 6553–6563, 1994.

[21] E. R. Wood, P. A. Dudchenko, and H. Eichenbaum, "The Global Record of Memory in Hippocampal Neuronal Activity," *Nature*, vol. 397, pp. 613–616, 1999.

[22] D. A. Redish, "The Hippocampal Debate : Are We Asking the Right Questions ?," *Behavioural brain research*, vol. 127, pp. 81–98, dec 2001.

[23] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, and E. I. Moser, "Microstructure of a Spatial Map in the Entorhinal Cortex," *Nature*, vol. 436, pp. 801–806, 2005.

[24] E. I. Moser, E. Kropff, and M.-B. Moser, "Place Cells, Grid Cells, and the Brain's Spatial Representation System," *Annual Review of Neuroscience*, vol. 31, pp. 69–89, jan 2008.

[25] J. B. Ranck, "Head Direction Cells in the Deep Cell Layer of Dorsolateral Pre-Subiculum in Freely Moving Rats," *Electrical Activity of the Archicortex*, 1985.

[26] J. S. Taube, R. U. Muller, and J. B. Ranck, "Head-Direction Cells Recorded From the Postsubiculum in Freely Moving Rats. I. Description and Quantitative Analysis," *Journal of Neuroscience*, vol. 10, no. 2, pp. 420–35, 1990.

[27] J. S. Taube and R. U. Muller, "Comparisons of Head Direction Cell Activity in the Postsubiculum and Anterior Thalamus of Freely Moving Rats," *Hippocampus*, vol. 8, pp. 87–108, jan 1998.

[28] E. Koechlin and Y. Burnod, "Dual Population Coding in the Neocortex : A Model of Interaction Between Representation and Attention in the Visual Cortex," *Journal of Cognitive Neuroscience*, vol. 8, no. 4, pp. 353–370, 1996.

[29] K. Gurney, T. J. Prescott, and P. Redgrave, "A Computational Model of Action Selection in the Basal Ganglia. II. Analysis and Simulation of Behaviour," *Biological cybernetics*, vol. 84, pp. 411–23, jun 2001.

[30] A. R. Cools, "Role of the Neostriatal Dopaminergic Activity in Sequencing and Selecting Behavioural Strategies : Facilitation of Processes Involved in Selecting the Best Strategy in a Stressful Situation," *Behavioural brain research*, vol. 1, pp. 361–378, 1980.

[31] J. W. Mink and W. T. Thach, "Basal Ganglia Intrinsic Circuits and Their Role in Behavior," *Current opinion in neurobiology*, vol. 3, pp. 950–7, dec 1993.

[32] J. D. Kropotov and S. C. Etlinger, "Selection of Actions in the Basal Ganglia-Thalamocortical Circuits : Review and Model," *International journal of psychophysiology : official journal of the International Organization of Psychophysiology*, vol. 31, pp. 197–217, mar 1999.

[33] T. J. Prescott, P. Redgrave, and K. Gurney, "Layered Control Architectures in Robots and Vertebrates," *Adaptive Behavior*, vol. 7, pp. 99–127, jan 1999.

[34] P. Redgrave, T. J. Prescott, and K. Gurney, "The Basal Ganglia : A Vertebrate Solution to the Selection Problem ?," *Neuroscience*, vol. 89, no. 4, pp. 1009–1023, 1999.

[35] J. S. Taube, "The Head Direction Signal : Origins and Sensory-Motor Integration," *Annual Review of Neuroscience*, vol. 30, pp. 181–207, jan 2007.

[36] A. A. Makarenko, S. B. Williams, F. Bourgault, and H. F. Durrant-whyte, "An Experiment in Integrated Exploration," in *IEEE International Conference on Robots and Systems*, 2002.

[37] C. Stachniss, H. Dirk, and W. Burgard, "Exploration with Active Loop-Closing for FastSLAM," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004)*, 2004.

[38] J.-A. Meyer and A. Guillot, "Simulation of Adaptive Behavior in Animats : Review and Prospect," in *Proceedings of the First International Conference on Simulation of Adaptive Behavior on From Animals to Animats*, pp. 2–14, MIT Press, 1991.

[39] H. Durrant-Whyte and T. Bailey, "Simultaneous Localization and Mapping (SLAM) : Part I," *IEEE Robotics Automation Magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[40] R. Smith, M. Self, and P. Cheeseman, "Estimating Uncertain Spatial Relationships in Robotics," *Proceedings. 1987 IEEE International Conference on Robotics and Automation*, vol. 4, no. January 1986, pp. 850–850, 1986.

[41] R. Smith, M. Self, and P. Cheeseman, "A Stochastic Map for Uncertain Spatial Relationships," *Proceedings of the 4th International Symposium on Robotics Research*, pp. 467–474, 1988.

[42] S. Thrun and J. J. Leonard, "Simultaneous Localization and Mapping," in *Springer Handbook of Robotics*, pp. 871–889, 2008.

[43] G. Henneberger, B. J. Brunsbach, and T. Klepsch, "Field Oriented Control of Synchronous and Asynchronous Drives Without Mechanical Sensors Using a Kalman-Filter," in *European Conference on Power Electronics and Applications*, vol. 3, p. 664, 1992.

[44] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "FastSLAM : A Factored Solution to the Simultaneous Localization and Mapping Problem," *Proc. of 8th National Conference on Artificial Intelligence/14th Conference on Innovative Applications of Artificial Intelligence*, vol. 68, no. 2, pp. 593–598, 2002.

[45] A. Doucet, N. D. Freitas, K. P. Murphy, and S. J. Russell, "Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks," *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pp. 176–183, 2000.

[46] G. Grisetti, C. Stachniss, and W. Burgard, "Improved Techniques for Grid Mapping with Rao-Blackwellized Particle Filters," *IEEE Transactions on Robotics*, vol. 23, no. 1, pp. 34–46, 2007.

[47] V. V. Hafner, "Cognitive Maps in Rats and Robots," *Adaptive Behavior*, vol. 13, no. 2, pp. 87–96, 2005.

[48] N. Cuperlier, M. Quoy, C. Giovannangeli, P. Gaussier, and P. Laroque, "Transition Cells for Navigation and Planning in an Unknown Environment," *From Animals to Animats*, vol. 4095, pp. 286–297, 2006.

[49] M. J. Milford, F. Wyeth, Gordon, and D. Prasser, "RatSLAM : A Hippocampal Model for Simultaneous Localization and Mapping," in *IEEE International Conference on Robotics and Automation*, no. May 2004, 2004.

[50] M. Milford, G. Wyeth, and D. Prasser, "RatSLAM on the Edge : Revealing a Coherent Representation from an Overloaded Rat Brain," *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4060–4065, oct 2006.

[51] J.-A. Meyer, A. Guillot, B. Girard, M. Khamassi, P. Pirim, and A. Berthoz, "The Psikharpax Project : Towards Building an Artificial Rat," *Robotics and Autonomous Systems*, vol. 50, pp. 211–223, mar 2005.

[52] K. Gurney, T. J. Prescott, and P. Redgrave, "A Computational Model of Action Selection in the Basal Ganglia. I. A New Functional Anatomy," *Biological cybernetics*, vol. 84, no. 6, pp. 401–410, 2001.

[53] K. Gurney, T. J. Prescott, and P. Redgrave, "A Computational Model of Action Selection in the Basal Ganglia. II. Analysis and Simulation of Behaviour.," *Biological cybernetics*, vol. 84, no. 6, pp. 411–423, 2001.

[54] D. E. Berlyne, *Structure and Direction in Thinking*. 1965.

[55] R. M. Ryan and E. L. Deci, "Intrinsic and Extrinsic Motivations : Classic Definitions and New Directions," *Contemporary Educational Psychology*, vol. 25, pp. 54–67, jan 2000.

[56] D. E. Berlyne, "Novelty and Curiosity as Determinants of Exploratory Behavior," *British Journal of Psychology*, vol. 41, no. 1-2, pp. 68–80, 1950.

[57] P.-Y. Oudeyer and F. Kaplan, "How Can We Define Intrinsic Motivation ?," in *Epigenetics Robotics : Modeling Cognitive Development in Robotic Systems*, 2008.

[58] A. Baranes and P.-y. Oudeyer, "R-IAC : Robust Intrinsically Motivated Active Learning," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 3, pp. 155–169, 2009.

[59] A. Baranes and P.-Y. Oudeyer, "Intrinsically Motivated Goal Exploration for Active Motor Learning in Robots : A Case Study," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pp. 1766–1773, 2010.

[60] X. Huang and J. Weng, "Novelty and Reinforcement Learning in the Value System of Developmental Robots.," 2002.

[61] N. Roy, A. Mccallum, and M. W. Com, "Toward Optimal Active Learning through Monte Carlo Estimation of Error Reduction," in *international conference on Machine Learning*, 2001.

[62] P. Capdepuy, D. Polani, and C. L. Nehaniv, "Maximization of Potential Information Flow as a Universal Utility for Collective Behaviour," in *IEEE Symposium on Artificial Life*, pp. 207–213, Ieee, Apr. 2007.

[63] J. Schmidhuber, "Curios Model-Building Control Systems," in *International Joint Conference on Neural Netwroks*, (Singapore), pp. 1458–1463, 1991.

[64] L. Macedo, "Modeling Forms of Surprise in Artificial Agents : Empirical and Theoretical Study of Surprise Functions," *Proceedings of the Cognitive Science Society*, vol. 26, 2001.

[65] L. Macedo and A. Cardoso, "The role of Surprise, Curiosity and Hunger on Exploration of Unknown Environments Populated with Entities," in *2005 Purtuguese Conference on Artificial Intelligence*, pp. 47–53, Ieee, dec 2005.

[66] A. G. Barto, S. Singh, and N. Chentanez, "Intrinsically Motivated Learning of Hierarchical Collections of Skills," in *International Conference on Development and Learning*, pp. 112–119, 2004.

[67] P.-y. Oudeyer and V. V. Hafner, "Intrinsic Motivation Systems for Autonomous Mental Development," in *IEEE Transactions on Evolutionary Computation*, vol. 2, pp. 265–286, 2007.

[68] W. G. Walter, "A Machine that Learns," *Scientific American*, pp. 60–63, 1951.

[69] W. G. Walter, "An Imitation of Life," 1950.

[70] V. Braitenberg, *Vehicles : Experiments in Synthetic Psychology*. MIT Press, 1986.

[71] R. Brooks, "Intelligence Without Representation," *Artificial Intelligence*, vol. 47, pp. 139–159, 1991.

[72] D. W. Batteau, "The Role of the Pinna in Human Localization," *Proceedings of the Royal Society B : Biological Sciences*, vol. 168, no. 1011, pp. 158–180, 1967.

[73] T. Willis, *Cerebri Anatome : Cui Accessit Nervorum Descriptio et Usus*. Martyn and Allestry, 1664.

[74] T. Willis, *The Remaining Medical Works of that Famous and Renowned Physician*. Dring, Harper & Leigh, 1681.

[75] C. A. Ferrier D Golz F and Y. G, "Discussion on the localization of function in the cortex cerebri," *Transactions of the International Medical Congress*, vol. 1, pp. 228–242, 1881.

[76] C. Alain, S. R. Arnott, S. Hevenor, S. Graham, and C. L. Grady, ""What" and "where" in the human auditory system.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, pp. 12301–6, oct 2001.

[77] J. C. Middlebrooks and D. M. Green, "Sound Localization by Human Listeners," *Annual Review of Psychology*, vol. 42, pp. 135–159, 1991.

[78] J. Blauert, "The Psychophysics of Human Sound Localization," 1997.

[79] L. Rayleigh, "On Our Perception of Sound Direction," *Philosophical Magazine*, vol. 13, no. 74, 1907.

[80] B. G. Shinn-Cunningham, S. Santarelli, and N. Kopco, "Tori of confusion : Binaural localization cues for sources within reach of a listener," *The Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1627–1636, 2000.

[81] A. J. King, J. W. H. Schnupp, and T. P. Doubell, "The shape of ears to come : Dynamic coding of auditory space," *Trends in Cognitive Sciences*, vol. 5, no. 6, pp. 261–270, 2001.

[82] P. T. Young, "The Role of Head Movements in Auditory Localization," *Journal of Experimental Psychology*, vol. 14, no. 2, pp. 95–124, 1931.

[83] H. Wallach, "On Sound Localization," *The Journal of the Acoustical Society of America (JASA)*, vol. 10, no. 1, p. 83, 1938.

[84] W. R. Thurlow, J. W. Mangels, and P. S. Runge, "Head Movements During Sound Localization," *The Journal of the Acoustical Society of America (JASA)*, vol. 42, no. 2, pp. 489–493, 1967.

[85] W. R. Thurlow and P. S. Runge, "Effect of Induced Head Movements on Localization of Direction of Sounds," *The Journal of the Acoustical Society of America (JASA)*, vol. 42, no. 2, pp. 480–488, 1967.

[86] S. Perrett and W. Noble, "The Contribution of Head Motion Cues to Localization of Low-Pass Noise," *Perception & psychophysics*, vol. 59, no. 7, pp. 1018–1026, 1997.

[87] S. Perrett and W. Noble, "The Effect of Head Rotations on Vertical Plane Sound Localization," *The Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2325–2332, 1997.

[88] C. Kim, R. Mason, and T. Brookes, "Head Movements Made by Listeners in Experimental and Real-Life Listening Activities," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 425–438, 2013.

[89] R. F. Lyon, "Machine hearing : An emerging field," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 131–136, 2010.

[90] S. Argentieri, A. Portello, M. Bernard, P. Danès, and B. Gas, *Binaural Systems in Robotics*, pp. 225–253. Berlin, Heidelberg : Springer Berlin Heidelberg, 2013.

[91] D. Pressnitzer and D. Gnansia, "Real-time auditory models," *International Computer Music Conference*, pp. 295–298, 2005.

[92] B. A. Olshausen and D. J. Field, "How Close Are We to Understanding V1 ?," *Neural Computation*, vol. 1699, pp. 1665–1699, 2005.

[93] M. Mishkin, L. G. Ungerleider, and K. A. Macko, "Object vision and spatial vision : two cortical pathways," *Trends in Neurosciences*, vol. 6, no. C, pp. 414–417, 1983.

[94] L. G. Ungerleider and J. V. Haxby, "'What' and 'where' in the human brain.," *Current opinion in neurobiology*, vol. 4, no. 2, pp. 157–65, 1994.

[95] M. Mahowald, "The silicon retina," *An Analog VLSI System for Stereoscopic Vision*, pp. 4–65, 1994.

[96] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128x128 120 dB 15mus Latency Asynchronous Temporal Contrast Vision Sensor," *IEEE journal of solid-state circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[97] P. Bardow, A. J. Davison, and S. Leutenegger, "Simultaneous Optical Flow and Intensity Estimation from an Event Camera," *Cvpr*, pp. 884–892, 2016.

[98] M. Ahissar and S. Hochstein, "The Reverse Hierarchy Theory of Visual Perceptual Learning," *Trends in Cognitive Sciences*, vol. 8, pp. 457–64, Oct. 2004.

[99] I. Nelken and M. Ahissar, "High-Level and Low-Level Processing in the Auditory System : The Role of Primary Auditory Cortex," *Dynamic of Speech Production and Perception*, pp. 5–12, 2006.

[100] S. Shamma, "On the Emergence and Awareness of Auditory Objects," *PLoS biology*, vol. 6, p. e155, jun 2008.

[101] M. Nahum, I. Nelken, and M. Ahissar, "Low-level information and high-level perception : the case of speech in noise.," *PLoS biology*, vol. 6, p. e126, May 2008.

[102] O. Collignon, P. Voss, M. Lassonde, and F. Lepore, "Cross-modal plasticity for the spatial processing of sounds in visually deprived subjects," *Experimental Brain Research*, vol. 192, no. 3, pp. 343–358, 2009.

[103] B. Röder, W. Teder-SaÈlejaÈrvi, A. Sterr, F. Rösler, S. A. Hillyard, and H. J. Neville, "Improved auditory spatial tuning in blind humans," *Nature*, vol. 400, no. 6740, pp. 162–166, 1999.

[104] P. Voss, M. Lassonde, F. Gougoux, M. Fortin, J.-P. Guillemot, and F. Lepore, "Early- and Late-Onset Blind Individuals Show Supra-Normal Auditory Abilities in Far-Space," *Canadian Field-Naturalist*, vol. 14, no. Current Biology, pp. 1734–1738, 2004.

[105] M. Gori, G. Sandini, C. Martinoli, and D. C. Burr, "Impairment of auditory spatial localization in congenitally blind human subjects," *Brain*, vol. 137, no. 1, pp. 288–293, 2014.

[106] M. A. Meredith and B. E. Stein, "Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration," *Journal of Neurophysiology*, vol. 56, no. 3, pp. 640–662, 1986.

[107] B. E. Stein, W. Jiang, and T. R. Stanford, "Multisensory integration in single neurons of the midbrain," *The handbook of multisensory processes*, vol. 15, pp. 243–264, 2004.

[108] A. K. Moschovakis, "The superior colliculus and eye movement control," *Current opinion in neurobiology*, vol. 6, no. 6, pp. 811–816, 1996.

[109] P. J. May, "The mammalian superior colliculus : laminar structure and connections," in *Progress in Brain Research*, pp. 321–378, 2006.

[110] K. E. Cullen, "The Vestibular System : Multimodal Integration and Encoding of Self-Motion for Motor Control," *Trends in Neurosciences*, vol. 35, no. 3, pp. 185–196, 2014.

[111] J. C. Hay, H. L. Pick, and K. Ikeda, "Visual capture produced by prism spectacles.," *Psychonomic science*, 1965.

[112] H. L. Pick, D. H. Warren, and J. C. Hay, "Sensory conflict in judgments of spatial direction," *Attention, Perception, & Psychophysics*, vol. 6, no. 4, pp. 203–205, 1969.

[113] J. W. Gebhard and G. H. Mowbray, "On discriminating the rate of visual flicker and auditory flutter," *The American journal of psychology*, vol. 72, no. 4, pp. 521–529, 1959.

[114] R. B. Welch and D. H. Warren, "Immediate perceptual response to intersensory discrepancy.," *Psychological bulletin*, vol. 88, no. 3, p. 638, 1980.

[115] R. Fendrich and P. M. Corballis, "The temporal cross-capture of audition and vision," *Perception & Psychophysics*, vol. 63, no. 4, pp. 719–725, 2001.

[116] C. R. Scheier, R. Nijhawan, and S. Shimojo, "Sound Alters Visual Temporal Resolution," in *Investigative Ophthalmology & Visual Science*, vol. 40, pp. S792—-S792, 1999.

[117] L. Shams, C. A. Y. Kamitani, S. Thompson, and S. Shimojo, "Sound Alters Visual Evoked Potentials in Humans," *Cognitive Neuroscience and Neuropsychology*, vol. 12, no. 17, pp. 3849–3852, 2001.

[118] L. Shams, Y. Kamitani, and S. Shimojo, "Visual Illusion Induced by Sound," *Cognitive Brain Research*, vol. 14, pp. 147–152, 2002.

[119] H. M. Saldana and L. D. Rosenblum, "Visual influences on auditory pluck and bow judgments," vol. 54, no. 3, pp. 406–416, 1993.

[120] D. A. Robinson, "Oculomotor Control Signals," *Basic mechanisms of ocular motility and their clinical implication*, pp. 337–374, 1975.

[121] K. Noda, H. Arie, Y. Suga, and T. Ogata, "Multimodal Integration Learning of Robot Behavior Using Deep Neural Networks," *Robotics and Autonomous Systems*, vol. 62, no. 6, pp. 721–736, 2014.

[122] J. Martens, "Deep Learning via Hessian-Free Optimization," in *international Conference on Machine Learning*, (Haifa, Israel), 2010.

[123] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, no. July 1928, pp. 379–423, 1948.

[124] W. James, *The Principles of Psychology*. Read Books Ltd, 1890.

[125] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A Coherent Computational Approach to Model Bottom-Up Visual Attention," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 802–817, 2006.

[126] J. Driver and C. Spence, "Attention and Cross Modal Construction of Space," *Trends in Cognitive Sciences*, vol. 2, no. 7, pp. 254–262, 1998.

[127] A. M. Treisman and G. Gelade, "A Feature-Integration Theory of Attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[128] M. I. Posner and S. E. Petersen, "The Attention System of the Human Brain," tech. rep., 1990.

[129] S. Petersen and M. Posner, "The Attention System of the Human Brain : 20 Years After," *Annual Review of Neuroscience*, vol. 21, no. 35, pp. 73–89, 2012.

[130] M. Corbetta and G. L. Shulman, "Control of Goal-Directed and Stimulus-Driven Attention in the Brain," *Nature Reviews Neuroscience*, vol. 3, no. 3, pp. 201–215, 2002.

[131] J. B. Hopfinger, M. H. Buonocore, and G. R. Mangun, "The Neural Mechanisms of Top-Down Attentional Control," *Nature neuroscience*, vol. 3, no. 3, pp. 284–291, 2000.

[132] M. Corbetta, J. M. Kincade, J. M. Ollinger, M. P. McAvoy, and G. L. Shulman, "Voluntary Orienting is Dissociated From Target Detection in Human Posterior Parietal Cortex," *Nature Neuroscience*, vol. 3, no. 3, pp. 292–297, 2000.

[133] K. S. LaBar, D. R. Gitelman, T. B. Parrish, and M. Mesulam, "Neuroanatomic Overlap of Working Memory and Spatial Attention Networks : A Functional MRI Comparison within Subjects," *Neuroimage*, vol. 10, no. 6, pp. 695–704, 1999.

[134] L. Pessoa, E. Gutierrez, P. A. Bandettini, and L. G. Ungerleider, "Neural Correlates of Visual Working Memory : fMRI Amplitude Predicts Task Performance," *Neuron*, vol. 35, no. 5, pp. 975–987, 2002.

[135] J. Downar, A. P. Crawley, D. J. Mikulis, and K. D. Davis, "A Multimodal Cortical Network for the Detection of Changes in the Sensory Environment," *Nature Neuroscience*, vol. 3, no. 3, pp. 277–283, 2000.

[136] J. M. Kincade, R. A. Abrams, S. V. Astafiev, G. L. Shulman, and M. Corbetta, "An Event-Related Functional Magnetic Resonance Imaging Study of Voluntary and Stimulus-Driven Orienting of Attention," *Journal of Neuroscience*, vol. 25, no. 18, pp. 4593–4604, 2005.

[137] I. Indovina and E. Macaluso, "Dissociation of Stimulus Relevance and Saliency Factors during Shifts of Visuospatial Attention," *Cerebral Cortex*, vol. 17, no. July, pp. 1701–1711, 2007.

[138] G. L. Shulman, M. P. Mcavoy, M. C. Cowan, S. V. Astafiev, A. P. Tansy, G. Avossa, M. Corbetta, L. Gordon, M. P. Mcavoy, M. C. Cowan, V. Astafiev, A. P. Tansy, and G. Avossa, "Quantitative Analysis of Attention and Detection Signals During Visual Search," pp. 3384–3397, 2003.

[139] J. J. Todd, D. Fougnie, and R. Marois, "Visual Short-Term Memory Load Suppresses Temporo-Parietal Junction Activity and Induces Inattentional Blindness," *Psychological Science*, vol. 16, no. 12, pp. 965–972, 2005.

[140] H.-C. Nothdurft, "Salience and Target Selection in Visual Search," *Visual Cognition*, vol. 14, no. 4-8, pp. 514–542, 2006.

[141] V. Duangudom and D. V. Anderson, "Using Auditory Saliency to Understand Complex Auditory Scenes," in *European Signal Processing Conference*, no. 15th, 2007.

[142] J. M. Wolfe, "Guided Search 2.0 - A Revised Model of Visual Search," *Psychonomic Bulletin & Review*, vol. 1, no. 2, pp. 202–238, 1994.

[143] M. A. Mcdaniel, M. J. Guynn, E. L. Glisky, S. R. Rubin, and B. C. Routhieaux, "Prospective Memory : A Neuropsychological Study," *Neuropsychology*, vol. 13, no. 1, pp. 103–110, 1999.

[144] Z. Li, "A Saliency Map in Primary Visual Cortex," *Trends in Cognitive Sciences*, vol. 6, no. 1, pp. 9–16, 2002.

[145] J. A. Mazer and J. L. Gallant, "Goal-Related Activity in V4 during Free Viewing Visual Search : Evidence for a Ventral Stream Visual Salience Map," *Neuron*, vol. 40, pp. 1241–1250, 2003.

[146] J. W. Bisley and M. E. Goldberg, "Neural Correlates of Attention and Distractibility in the Lateral Intraparietal Area," *Journal of Neurophysology*, vol. 95, pp. 1696–1717, 2006.

[147] K. G. Thompson and N. P. Bichot, "A Visual Salience Map in the Primate Frontal Eye Field," *Progress in Brain Research*, vol. 147, 2005.

[148] W. A. Yost, "Auditory Perception and Sound Source Determination," *Current Directions in Psychological Science*, vol. 1, no. 6, pp. 179–184, 1992.

[149] J. W. Hall, M. P. Haggard, and M. A. Fernandes, "Detection in noise by spectro-temporal pattern analysis," *The Journal of the Acoustical Society of America*, vol. 76, no. 1, pp. 50–56, 1984.

[150] S. Onat, K. Libertus, and P. König, "Integrating Audiovisual Information for the Control of Overt Attention," *Journal of Vision*, vol. 7, no. 10, p. 11, 2007.

[151] T. R. Stanford, "Evaluating the Operations Underlying Multisensory Integration in the Cat Superior Colliculus," *Journal of Neuroscience*, vol. 25, no. 28, pp. 6499–6508, 2005.

[152] I. P. Howard and W. B. Templeton, "Human Spatial Orientation," 1966.

[153] J. J. Gibson, "Adaptation, After-Effect and Contrast in the Perception of Curved Lines," *Journal of Experimental Psychology*, vol. 16, no. 1, pp. 1–31, 1933.

[154] F. B. Colavita, "Human Sensory Dominance," *Attention, Perception, & Psychophysics*, vol. 16, no. 2, pp. 409–412, 1974.

[155] M. I. Posner, M. J. Nissen, and R. M. Klein, "Visual Dominance : An Information-Processing Account of its Origins and Significance," *Psychological Review*, vol. 83, no. 2, pp. 157–171, 1976.

[156] M. Turatto, F. Benso, G. Galfano, and C. Umiltà, "Nonspatial Attentional Shifts Between Audition and Vision," *Journal of Experimental Psychology : Human Perception and Performance*, vol. 28, no. 3, pp. 628–639, 2002.

[157] C. J. Spence and J. Driver, "Covert Spatial Orienting in Audition : Exogenous and Endogenous Mechanisms," *Journal of Experimental Psychology : Human Perception and Performance*, vol. 20, no. 3, p. 555, 1994.

[158] C. Spence and J. Driver, "Audiovisual Links in Endogenous Covert Spatial Attention," *Journal of Experimental Psychology : Human Perception and Performance*, vol. 22, no. 4, p. 1005, 1996.

[159] C. Spence and J. Driver, "Audiovisual Links in Exogenous Covert Spatial Orienting," *Perception & Psychophysics*, vol. 59, no. 1, pp. 1–22, 1997.

[160] C. Spence and J. Driver, "On Measuring Selective Attention to an Expected Sensory Modality," *Perception & Psychophysics*, vol. 59, no. 3, pp. 389–403, 1997.

[161] A. Berti and F. Frassinetti, "When Far Becomes Near : Remapping of Space," *Journal of Cognitive Neuroscience*, vol. 12, no. 3, pp. 415–420, 2000.

[162] T. Hosoya, S. A. Baccus, and M. Meister, "Dynamic predictive coding by the retina," *Nature*, vol. 436, pp. 71–77, jul 2005.

[163] K. Friston, "A Theory of Cortical Responses," *Philosophical Transactions : Biological Sciences*, vol. 360, no. 1456, pp. 815–836, 2005.

[164] T. Lochmann and S. Deneve, "Neural processing as causal inference," *Current Opinion in Neurobiology*, vol. 21, no. 5, pp. 774–781, 2011.

[165] L. H. Arnal and A.-L. Giraud, "Cortical oscillations and sensory predictions.," *Trends in cognitive sciences*, vol. 16, pp. 390–8, July 2012.

[166] M. Sams, P. Paavilainen, K. Alho, and R. Näätänen, "Auditory Frequency Discrimination and Event-Related Potentials," *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, vol. 62, no. 6, pp. 437–448, 1985.

[167] R. Näätänen, P. Paavilainen, T. Rinne, and K. Alho, "The mismatch negativity (MMN) in basic research of central auditory processing : a review.," *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, vol. 118, pp. 2544–90, Dec. 2007.

[168] R. Näätänen, A. Gaillard, and S. Mäntysalo, "Early Selective-Attention Effect on Evoked Potential Reinterpreted," *Acta Psychologica*, vol. 42, pp. 313–329, 1978.

[169] S. Molholm, A. Martinez, W. Ritter, D. C. Javitt, and J. J. Foxe, "The neural circuitry of pre-attentive auditory change-detection : An fMRI study of pitch and duration mismatch negativity generators," *Cerebral Cortex*, vol. 15, no. 5, pp. 545–551, 2005.

[170] K. Alho, "Cerebral Generators of Mismatch Negativity (MMN) and Its Magnetic Counterpart (MMNm) Elicited by Sound Changes," *Ear and Hearing*, vol. 16, no. 1, pp. 38–51, 1995.

[171] R. Näätänen and K. Alho, "Generators of Electrical and Magnetic Mismatch Responses in Humans," *Brain topography*, vol. 7, no. 4, pp. 315–320, 1995.

[172] J. A. Winer and C. E. Schreiner, eds., *The Auditory Cortex*. Springer.

[173] C. Koch and S. Ullman, "Shifts in Selective Visual Attention : Towards the Underlying Nneural Circuitry," *Human neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[174] L. Itti and C. Koch, "Computational Modelling of Visual Attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[175] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[176] P. Reinagel and A. M. Zador, "The Effect of Gaze on Natural Scene Statistics," in *Neural Information and Coding Workshop*, pp. 16–20, 1997.

[177] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-Down Control of Visual Attention in Object Detection," *IEEE International Conference on Image Processing, September 14-17*, vol. 1, pp. 1–4, 2003.

[178] E. Simoncelli and W. Freeman, "The Steerable Pyramid : A Flexible Architecture for Multi-Scale Derivative Computation," in *International Conference on Image Processing*, vol. 3, (Washington, DC), pp. 444–447, 1995.

[179] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for Allocating Auditory Attention : An Auditory Saliency Map," *Current Biology*, vol. 15, pp. 1943–1947, 2005.

[180] O. Kalinli and S. Narayanan, "A Saliency-Based Auditory Attention Model with Applications to Unsupervised Prominent Syllable Detection in Speech," in *Interspeech*, pp. 1–4, 2007.

[181] O. Kalinli, S. Sundaram, and S. Narayanan, "Saliency-Driven Unstructured Acoustic Scene Classification Using Latent Perceptual Indexing," in *Multimedia Signal Processing*, (Rio de Janeiro), 2009.

[182] S. Sundaram and S. Narayanan, "Audio Retrieval by Latent Perceptual Indexing," in *Internation Conference on Acoustic, Speech and Signal Processing - ICASSP*, (Las Vega, NV, USA), pp. 769–772, 2008.

[183] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational Auditory Scene Recognition," *IEEE International Conference on Audio, Speech and Signal Processing*, pp. II–1941–II–1944, 2002.

[184] S. Deerwester, S. T. Dumais, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[185] J. Ruesch, M. Lopes, A. Bernardino, J. Hörnstein, J. Santos-Victor, and R. Pfeifer, "Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 962–967, 2008.

[186] V. Tikhanoff, A. Cangelosi, P. Fitzpatrick, G. Metta, L. Natale, and F. Nori, "An Open-Source Simulator for Cognitive Robotics Research : The Prototype of the iCub Humanoid Robot Simulator," *PerMIS '08 Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, pp. 57–61, 2008.

[187] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor, A. Bernardino, and L. Montesano, "The iCub Humanoid Robot : An Open-Systems Platform for Research in Cognitive Development," *Neural Networks*, vol. 23, no. 8-9, pp. 1125–1134, 2010.

[188] R. A. Peters II, K. E. Hambuchen, K. Kawamura, and D. M. Wilkes, "The Sensory Ego-Sphere as a Short-Term Memory for Humanoids," *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, no. 1, pp. 451–459, 2001.

[189] S. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised Machine Learning : a Review of Classification Techniques," *Informatica*, vol. 31, pp. 249–268, 2007.

[190] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour Detection and Hierarchical Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 898–916, 2011.

[191] J. Deng, W. Dong, R. Socher, L.-j. Li, K. Li, and L. Fei-fei, "ImageNet : A Large-Scale Hierarchical Image Database," in *Computer Vision and Pattern Recognition 2009*, pp. 2–9, 2009.

[192] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances In Neural Information Processing Systems*, pp. 1–9, 2012.

[193] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[194] Y. Ephraim, "A Bayesian Estimation Approach for Speech Enhancement Using Hidden Markov Models," 1992.

[195] T. V. Sreenivas and P. Kirnapure, "Codebook Constrained Wiener Filtering for Speech Enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 383–389, 1996.

[196] H. Veisi and H. Sameti, "Speech Enhancement Using Hidden Markov Models in Mel-Frequency Domain," *Speech Communication*, vol. 55, no. 2, pp. 205–220, 2013.

[197] S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[198] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[199] R. Martin, "Speech Enhancement Based on Minimum Mean-Square Error Estimation and Supergaussian Priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, 2005.

[200] A. Harma, M. F. Mckinney, and J. Skowronek, "Automatic Surveillance of the Acoustic Activity in our Living Environment," in *International Conference on Multimedia and Expo*, vol. 1, pp. 6–9, 2005.

[201] P. Dangauthier, P. Bessiere, A. Spalanzani, P. Dangauthier, P. Bessiere, A. Spalanzani, P. Dangauthier, and P. Bessi, "Auto-Supervised Learning in the Bayesian Programming Framework," in *International Conference on Robotics and Automation*, pp. 1–6, 2005.

[202] J. Weng, J. Mcclelland, A. Pentland, O. Sporns, M. Sur, and E. Thelen, "Autonomous Mental Development by Robots and Animals," *Science*, vol. 291, no. 5504, pp. 599–600, 2001.

[203] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement Learning : A Survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.

[204] G.-b. Huang, S. Member, Q.-y. Zhu, and C.-k. Siew, "Real-Time Learning Capability of Neural Networks," vol. 17, no. 4, pp. 863–878, 2006.

[205] T. Kohonen, "Self-Organized Formation of Popologically Correct Feature Maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[206] T. Kohonen, "Essentials of the Self-Organizing Map," *Neural Networks*, vol. 37, pp. 52–65, 2013.

[207] T. Kohonen, "Clustering, Taxonomy, and Topological Maps of Patterns," in *Proc. 6th ICPR, Int. Conf. on Pattern Recognition*, (Washington, DC), pp. 114–128, IEEE Computer Soc. Press, 1982.

[208] T. Kohonen, "The Self-Organizing Map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[209] S. Kaski, J. Kangas, and T. Kohonen, "Bibliography of Self-Organizing Map (SOM) Papers : 1981-1997," *Neural Computing Surveys*, pp. 102–350, 1998.

[210] M. Oja, S. Kaski, and T. Kohonen, "Bibliography of Self-Organizing Map (SOM) Papers : 1998-2001 Addendum," tech. rep., 2003.

[211] M. Polla, T. Honkela, and T. Kohonen, "Bibliography of Self-Organizing Map (SOM) Papers : 2002–2005 Addendum," tech. rep., 2009.

[212] D. Ruta and B. Gabrys, "An Overview of Classifier Fusion Methods," *Computing and Information Systems*, vol. 7, no. 1, pp. 1–10, 2000.

[213] D. L. Hall and J. Llinas, "An Introduction to Multisensor Data Fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[214] A. J. C. Sharkey, *Combining Artificial Neural Nets : Ensemble and Modular Multi-Net Systems (Perspectives in Neural Computing)*. Springer Verlag, 1998.

[215] F. M. Alkoot and J. Kittler, "Experimental Evaluation of Expert Fusion Strategies," *Pattern Recognition Letters*, vol. 20, no. 11–13, pp. 1361–1369, 1999.

[216] L. I. Kuncheva, "A Theoretical Study on Six Classifier Fusion Strategies," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 281–286, 2002.

[217] X. Zou and B. Bir, "Tracking Humans using Multi-modal Fusion," *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, p. 4, 2005.

[218] R. C. Luo and C.-C. Yih, "Multisensor Fusion and Integration : Approaches, Applications and Future Research Directions," 2002.

[219] P. S. Aleksic and A. K. Katsaggelos, "Audio-Visual Biometrics," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 2025–2044, 2006.

[220] A. Jaimes and N. Sebe, "Multimodal Human-Computer Interaction : A Survey," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 116–134, 2007.

[221] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, *Multimodal Fusion for Multimedia Analysis : A Survey*, vol. 16. 2010.

[222] Z. H. Zhou, J. Wu, and W. Tang, "Ensembling Neural Networks : Many Could Be Better Than All," *Artificial Intelligence*, vol. 137, no. 1-2, pp. 239–263, 2002.

[223] Z. Wu, L. Cai, and H. Meng, "Multi-level Fusion of Audio and Visual Features for Speaker Identification," *Advances in Biometrics*, pp. 493–499, 2005.

[224] G. J. Klir and T. A. Folger, "Fuzzy Sets, Uncertainty, and Information," 1988.

[225] C.-l. Benjamin, A. Sylvain, and G. Bruno, "Multimodal Fusion and Inference Using Binaural Audition and Vision," in *International Congress on Acoustics*, 2016.

[226] D. D. Corkill, "Blackboard systems," *Artificial Intelligence Review*, vol. 2, no. 2, pp. 103–118, 1991.

[227] J. Blauert, "Binaural models and their technological applications," in *International Congress on Sound and Vibration*, (Vilnius, Lithuania), pp. 1–4, 2012.

[228] J. Blauert, "A perceptionist's view on psychoacoustics," *Archives of Acoustics*, vol. 37, no. 3, pp. 365–371, 2012.

[229] D. Calisi, A. Farinelli, L. Locci, and D. Nardi, "Multi-objective Exploration and Search for Autonomous Rescue Robots," *Journal of Field Robotics*, vol. 24, no. 8/9, pp. 763–777, 2007.

[230] TwoEars, "Full proposal ICT FET Open Call - Two !Ears," tech. rep., 2013.

[231] L. D. Erman, F. Hayes-Roth, V. R. Lesser, and D. R. Reddy, "The Hearsay-II Speech-Understanding System : Integrating Knowledge to Resolve Uncertainty," *ACM Computing Surveys*, vol. 12, no. 2, pp. 213–253, 1980.

[232] N. Ma, G. Brown, and T. May, *Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions.* 2015.

[233] F. L. Wightman and D. J. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement.," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2841–2853, 1999.

[234] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, pp. 1–13, Jan 2011.

[235] J. Woodruff, S. Member, and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments Binaural localization of multiple sources in reverberant and noisy environments," vol. 20, no. c, pp. 1503–1512, 2012.

[236] N. Ma, I. Trowitzsch, Y. Kashef, J. Mohr, K. Obermayer, C. Schymura, D. Kolossa, T. Walther, H. Wierstorf, T. May, G. Brown, B. Cohen-Lhyver, P. Danès, M. Devy, T. Forgue, A. Podlubne, and B. Vandeportaele, "Report on Evaluation of the Two !Ears Expert System," tech. rep., 2012.

[237] D. Wang and G. J. Brown, *Computational auditory scene analysis : Principles, algorithms, and applications.* Wiley-IEEE Press, 2006.

[238] A. Bregman, "Auditory Scene Analysis," in *International Encyclopedia of the Social and Behavioral Sciences*, 1990.

[239] R. Sutton, "Introduction to Reinforcement Learning,"

[240] P. Reverdy and N. E. Leonard, "Parameter Estimation in Softmax decision-making models with linear objective functions," *EEE Transactions on Automation Science and Engineering*, vol. 13, no. 1, pp. 54–67, 2016.

[241] T. May, S. Van De Par, and A. Kohlrausch, "Binaural localization and detection of speakers in complex acoustic scenes," in *The Technology of Binaural Listening*, pp. 397–425, Springer, 2013.

[242] T. May, N. Ma, and G. J. Brown, "Robust Localization of Multiple Speakers Exploiting Head Movements and Multi-Conditional Training of Binaural Cues," *Internation Conference on Acoustic, Speech and Signal Processing - ICASSP*, pp. 2679–2683, 2015.

[243] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[244] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[245] R. D. Patterson and J. Holdsworth, "A functional model of neural activity patterns and auditory images," *Advances in Speech, Hearing, and Language Processing*, vol. 3, pp. 547–563, 1996.

[246] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, 2001.

[247] T. May, S. Van De Par, and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 7, pp. 2016–2030, 2012.

[248] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[249] K. Jensen and T. H. Andersen, "Real-time beat estimationusing feature extraction," in *International Symposium on Computer Music Modeling and Retrieval*, pp. 13–22, Springer, 2003.

[250] H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky, "Spectral entropy based feature for robust ASR," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1, pp. I—-193, IEEE, 2004.

[251] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The timbre toolbox : Extracting audio descriptors from musical signals," *The Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2902–2916, 2011.

[252] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 6, pp. 3089–3092, IEEE, 1999.

[253] N. Moritz, J. Anemüller, and B. Kollmeier, "Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 5492–5495, IEEE, 2011.

[254] T. May and T. Dau, "Computational speech segregation based on an auditory-inspired modulation analysis," *The Journal of the Acoustical Society of America*, vol. 136, no. 6, pp. 3350–3359, 2014.

[255] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. 4134–4151, 2012.

[256] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The Balanced Accuracy and its Posterior Distribution," *Proceedings - International Conference on Pattern Recognition*, pp. 3121–3124, 2010.

[257] N. Ma, I. Trowitzsch, Y. Kashef, J. Mohr, K. Obermayer, C. Schymura, D. Kolossa, T. Walther, H. Wierstorf, T. May, G. Brown, B. Cohen-Lhyver, P. Danès, M. Devy, T. Forgue, A. Podlubne, and B. Vandeportaele, "Report on Evaluation of the TwoEars Expert System," tech. rep., 2016.

[258] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit, "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 858–865, 2011.

[259] S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, N. Navab, P. Fua, and V. Lepetit, "Gradient Response Maps for Real-Time Detection of Textureless Objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 1–14, 2012.

[260] S. Argentieri, G. Bustamante, B. Cohen-Lhyver, P. Danès, X. Dollat, T. Forgue, B. Gas, M. Herrb, A. Mallet, J. Manhès, A. Musabini, J. Piat, A. Podlubne, and V. Bertrand, "Deliverable 5.3 : Final Report on Hardware / Software Integration and Robotics Test Bed," tech. rep., 2016.

[261] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME Speech Separation and Recognition Challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, 2013.

[262] B. Cohen-Lhyver, S. Argentieri, and B. Gas, "Modulating the Auditory Turn-to Reflex on the Basis of Multimodal Feedback Loops : the Dynamic Weighting Model," in *IEEE Robio*, 2015.

[263] T. Walther and B. Cohen-Lhyver, "Multimodal Feedback in Auditory-Based Active Scene Exploration," in *Forum Acusticum*, 2014.

[264] B. Girard, V. Cuzin, A. Guillot, K. N. Gurney, and T. J. Prescott, "Comparing a Brain-Inspired Robot Action Selection Mechanism With 'Winner-Takes-All'," in *From Animals to Animats 7 : Proceedings of the seventh international conference on simulation of adaptive behavior*, vol. 7, p. 75, MIT Press, 2002.

[265] T. Kohonen, *Self-Organizing Maps*. Berlin, Germany : Springer-Verlag Berlin Heidelberg, 2001.

[266] A. Ultsch, "Self Organized Feature Maps for Monitoring and Knowledge Aquisition of a Chemical Process," in *ICANN : International Conference on Artificial Neural Networks*, pp. 864–867, Springer, 1993.

[267] A. Ultsch, "U *-Matrix : A Tool to Visualize Clusters in High Dimensional Data," *Computer*, vol. 52, no. 36, pp. 1–12, 2003.

[268] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Human Genetics*, 1936.

[269] R. Tatoian and L. Hamel, "Self-Organizing Map Convergence Self-Organizing Maps," in *International Conference on Data Mining*, p. 92, 2016.

[270] H. Yin and N. M. Allinson, "On the Distribution and Convergence of Feature Space in Self-Organizing Maps," *Neural computation*, vol. 7, no. 6, pp. 1178–1187, 1995.