



HAL
open science

L'interaction humain-robot multimodale proactive dans un cadre hospitalier

Timothée Dhaussy

► **To cite this version:**

Timothée Dhaussy. L'interaction humain-robot multimodale proactive dans un cadre hospitalier. Autre [cs.OH]. Université d'Avignon, 2024. Français. NNT : 2024AVIG0106 . tel-04953930

HAL Id: tel-04953930

<https://theses.hal.science/tel-04953930v1>

Submitted on 18 Feb 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



THÈSE DE DOCTORAT D'AVIGNON UNIVERSITÉ

École Doctorale n°536
Agrosciences et Sciences

Mention de doctorat : Informatique

Laboratoire : Laboratoire Informatique d'Avignon

L'interaction humain-robot multimodale proactive dans un cadre hospitalier

Présentée par
Timothée Dhaussy

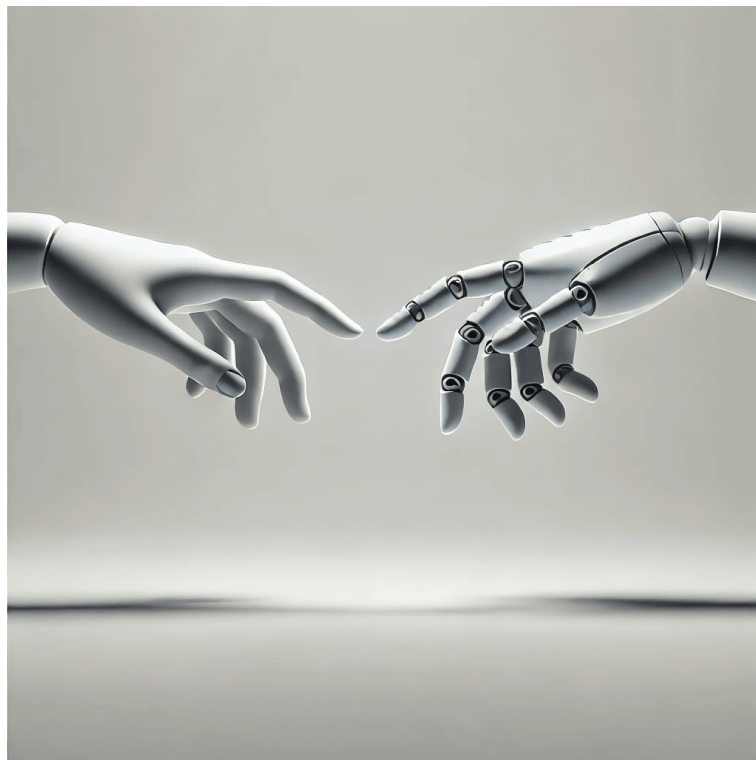
Soutenue publiquement le 21 octobre 2024 devant le jury composé de :

Dr Aurélie Clodic	LAAS-CNRS	Rapporteuse
Pr Julien Pinquier	Université de Toulouse, IRIT	Rapporteur
Pr Laurence Devillers	Sorbonne Université, LISN-CNRS	Examinatrice
Pr Olivier Alata	Université Jean Monnet, Laboratoire Hubert Curien	Examineur
Pr Fabrice Lefèvre	Avignon Université, LIA	Directeur de thèse
Dr Bassam Jabaian	Avignon Université, LIA	Co-encadrant



L'INTERACTION HUMAIN-ROBOT MULTIMODAL PROACTIVE
DANS UN CADRE HOSPITALIER

TIMOTHÉE DHAUSSY



Septembre 2024

À ma grand mère Michelle.

ABSTRACT

L'interaction humain-robot (Human-Robot Interaction ou HRI) est un domaine de recherche interdisciplinaire en robotique et en sciences sociales. Elle vise à comprendre, concevoir et évaluer l'utilisation des robots par les humains. Elle représente un enjeu important de la robotique sociale au 21^{ème} siècle. Ces derniers temps, l'intérêt pour les robots compagnons aptes à aider les individus dans leur quotidien et à interagir avec eux a considérablement augmenté. Ces robots, considérés comme des entités sociales, ont démontré leur utilité dans les domaines des soins de santé et du bien-être psychologique des personnes âgées. La proactivité, ou la capacité à agir de manière anticipée et autonome, est une caractéristique intrinsèquement humaine qui permet d'influencer activement son environnement et ses circonstances, plutôt que d'y réagir passivement. Dans les interactions humain-robot, cette proactivité est cruciale car elle permet aux robots de réagir de manière plus naturelle et adaptée aux besoins des utilisateurs. Pour anticiper les besoins des utilisateurs et prendre des initiatives, le robot doit comprendre son environnement à l'aide de ses perceptions multimodales.

Dans cette thèse, nous nous concentrons sur la création d'un système multimodal proactif pour le robot social Pepper, destiné à une salle d'attente d'hôpital. Pour ce faire, nous avons développé une architecture cognitive d'interaction humain-robot, fondée sur une boucle continue de perceptions, de représentation et de décision. Le flux de perceptions se divise en deux étapes : d'abord, la récupération des données des capteurs du robot, puis leur enrichissement grâce à des modules de raffinement. Un module de raffinement de diarisation du locuteur, basé sur une modélisation bayésienne de la fusion des perceptions audio et visuelles par coïncidence spatiale, a été intégré. Pour permettre une action proactive, nous avons conçu un modèle analysant la disponibilité des utilisateurs à une interaction dans une salle d'attente.

Les perceptions ainsi raffinées sont ensuite ordonnées et alignées pour créer une représentation de l'environnement constamment mise à jour. Cette image de l'environnement est alors transmise à la couche décisionnelle. Là, un module de planification d'action analyse les données environnementales et élabore des stratégies d'action en informant les modules d'action de manière asynchrone. Cette capacité à fonctionner de manière asynchrone permet au planificateur d'action de continuer à guetter les opportunités proactives que lui confère la scène, malgré le fonctionnement d'un des sous modules d'action tel que le module de parole, chargé de tenir une conversation avec un utilisateur lors d'une interaction. L'ensemble du système est implémenté sur Robot Operating System (ROS), permettant ainsi son adaptation à divers supports robotiques.

Cette thèse présente les mécanismes nécessaires à la création d'un système d'interaction multimodale proactif entre l'humain et le robot. Ce système inclut tous les modules de perception et d'action, ainsi qu'une architecture cognitive globale de la gestion des perceptions. L'ensemble a été testé dans un environnement de laboratoire contrôlé, ainsi que dans des conditions réelles à l'hôpital de Broca.

Mots clés : Interaction humain-robot, HRI, robotique sociale, proactivité, perception multimodale, architecture cognitive, Pepper, ROS, diarisation du locuteur,

REMERCIEMENTS

Je souhaite tout d'abord exprimer ma plus sincère reconnaissance envers mes directeurs de thèse, Fabrice Lefèvre, Radu Horaud et Bassam Jabaian, pour leur encadrement et leur soutien constants. Mes remerciements vont également aux membres du LIA, Teva pour son soutien sur Pepper, Driss et Pascal pour ses 3 ans de collaboration en base de programmation, Tania, les Lucas, Jarod, Mathias, Sondes, Virgile, Ahmed...

Bien que cela occupe une part importante, la vie d'un doctorant ne se limite pas à la recherche. Je suis immensément reconnaissant envers les personnes qui m'ont accompagné tout au long de cette thèse. Je tiens également à remercier chaleureusement mes amis de Mudaison et d'Avignon pour les moments merveilleux passés ensemble. Une pensée particulière à la team Vezzo, PG, Arthur, Luis, Thibault, Aph, Al, le T-crew, Tommy, Tommy, Tony, l'équipe Multiprise, Alex, Fabien et Mélissa et mes amis d'enfances de Mudaison, Georgette, Nono, Léo et Vico.

Enfin, je suis profondément reconnaissant envers mes parents, mon frère et ma sœur ainsi que mes grands parents pour leur amour inconditionnel et leur soutien indéfectible tout au long de mon parcours académique, ainsi que ma tante pour ses messages d'encouragement.

TABLE DES MATIÈRES

1	INTRODUCTION	3	
1.1	Motivation	3	
1.2	Définition des concepts	5	
1.3	Objectifs, le contexte du projet ANR μ DialBot	6	
1.4	Contributions	9	
1.5	Organisation des sections	9	
1.6	Publications	10	
2	TRAVAUX CONNEXES : INTERACTION HUMAIN-ROBOT ET PROACTIVITÉ		15
2.1	Introduction	15	
2.2	L'interaction humain-robot	15	
2.2.1	Définition de la Human Robot Interaction (HRI)	15	
2.2.2	Le robot social	16	
2.2.3	Les applications en HRI	19	
2.2.4	Les modèles HRI	21	
2.2.5	ROS	22	
2.2.6	Pepper un robot multimodal	24	
2.3	L'apprentissage machine	26	
2.3.1	Introduction	26	
2.3.2	Les approches non neuronales	27	
2.3.3	Les mélanges de lois de gaussiennes	27	
2.3.4	Les approches par réseaux de neurones	29	
2.4	Les signaux de l'environnement	31	
2.4.1	Reconnaissance automatique de la parole	31	
2.4.2	Détection et suivi de personnes par vision	32	
2.4.3	Diarisation du locuteur	32	
2.4.4	Les signaux sociaux	34	
2.5	La proactivité	36	
2.5.1	Dans la psychologie organisationnelle	36	
2.5.2	Dans la HRI	36	
2.5.3	Modélisation HRI de la pro-activité	37	
2.6	Projet μ DialBot	39	
2.6.1	Présentation	39	
2.6.2	Organisation	40	
2.6.3	Impact et bénéfice du projet	42	
3	CONTRIBUTIONS	45	
3.1	Introduction	45	
3.1.1	Contexte	45	
3.1.2	Introduction à l'architecture FlowAct	45	
3.2	Le flot de perception	46	
3.2.1	Extraction des perceptions brutes	46	
3.3	Raffinement des perceptions	48	

3.3.1	Introduction	48
3.3.2	Parole	49
3.3.3	Vision	50
3.3.4	Diarisation multimodale du locuteur	53
3.3.5	Analyse de l'engagement en HRI	65
3.3.6	Analyse de l'acceptation de l'engagement	71
3.3.7	Résumé intermédiaire	78
3.4	FlowAct, une architecture proactive d'interaction humain-robot	79
3.4.1	Fonctionnement général	79
3.4.2	Perception	81
3.4.3	La modélisation de l'environnement	83
3.4.4	Module de décision	86
3.4.5	Résumé intermédiaire	89
3.5	Expérimentations	89
3.5.1	Introduction	89
3.5.2	Expérience de laboratoire contrôlé	90
3.5.3	Expérience à l'hôpital Broca d'AP-HP	96
3.6	Conclusion	103
4	CONCLUSION, DISCUSSION ET PERSPECTIVES	107
4.1	Conclusion	107
4.2	Contributions	107
4.3	Discussions	108
4.4	Limitations	109
4.5	Perspectives	110
	BIBLIOGRAPHIE	113

TABLE DES FIGURES

Figure 1	Image de la salle d'expérience de l'hôpital Broca.	8
Figure 2	Diagramme des champs des disciplines impliquées dans l'Interaction Humain-Robot (GOLAM RASHED 2016).	16
Figure 3	Diagramme de l'espace conceptuel des approches HRI d'après (DAUTENHAHN 2007) A, socialement évocateur; B, socialement situé; C, sociable; D, socialement intelligent; E, socialement interactif.	18
Figure 4	Représentation de la vallée de l'étrange de MORI, MACDORMAN et KAGEKI (2012).	19
Figure 5	Exemples de robots sociaux, Quori (SPECIAN et al. 2022), Ari (COOPER et al. 2020), Moxie (HURST et al. 2020), Miroki (TOOLS 2023), Nao (JANISZEK et al. 2011), Jibo (RANE, MHATRE et KURUP 2014).	20
Figure 6	Diagramme de l'architecture de la communication dans le système ROS.	22
Figure 7	Représentation des capteurs du robot Pepper.	25
Figure 8	Diagramme de l'organisation des lots de travail chargés de la conception du système de HRI dans projet μ DialBot.	41
Figure 9	Diagramme de l'architecture FlowAct, une boucle continue de perceptions et d'actions.	47
Figure 10	Diagramme en bloc du module de détection de personne.	51
Figure 11	Diagramme en bloc de la méthode de diarisation du locuteur proposée.	54
Figure 12	Exemple de diarisation du locuteur sur deux participant du corpus Pepper, le cercle représente le locuteur dominant prédit.	61
Figure 13	Exemples de diarisation du locuteur sur trois participants, la vérité de base est représentée par un carré, la prédiction par un cercle, y_{tn}^* est le point rouge et les autres points verts plus petits sont les localisations sonores. Notre modèle fait la bonne prédiction sur les images 1-5, mais échoue sur les images 6-9 à prédire le locuteur précédent.	64
Figure 14	Exemples de l'ensemble de données Daisee, en haut, et UE-HRI, en bas.	67
Figure 15	Exemples d'interactions avec un Sign of Engagement Decrease (SED) détecté (pointillé bleu) et la prédiction d'engagement du modèle DAISEE (ligne rouge).	70
Figure 16	Diagramme de la modélisation de l'Interaction Acceptance Belief (IAB).	72
Figure 17	Exemples de comportements issus du jeu de données.	75
Figure 18	Diagramme des cellules Gate Recurrent Unit (GRU) en configuration 'many-to-one'.	76

Figure 19	Diagramme de l'architecture FlowAct, une boucle continue de perception et d'action.	80
Figure 20	Diagramme de la couche de perception de FlowAct.	82
Figure 21	Diagramme de la couche de représentation du système FlowAct.	83
Figure 22	Diagramme flowchart du fonctionnement du module de ré-identification de personne.	85
Figure 23	Diagramme de la couche de décision de FlowAct.	86
Figure 24	Diagramme du schéma expérimentale de l'architecture FlowAct.	91
Figure 25	Diagrammes en boîte des résultats du questionnaire Usability Metric for User Experience (UMUX).	94
Figure 26	Exemple d'une scène expérimentale, montrant l'IAB pour chaque personne détectée, avec son id. Les contours rouge indiquent une personne disponible à l'engagement.	95
Figure 27	Questionnaire d'expérience utilisateur Acceptability E-Scale (AES).	96
Figure 28	Questionnaire d'expérience utilisateur System Usability Scale (SUS).	97
Figure 29	Diagrammes en boîte des résultats du questionnaire SUS.	99
Figure 30	Diagrammes en boîte des résultats du questionnaire AES.	99
Figure 31	Exemple d'une séquence d'engagements successif.	101
Figure 32	Exemple de conversation entre un patient et le robot Pepper.	102

LISTE DES TABLEAUX

Table 1	Tableau récapitulatif des définitions des termes utilisés dans l'étude	6
Table 2	Tableau des indices comportementaux associés à certains des comportements sociaux les plus importants VINCIARELLI, Maja PANTIC et al. 2008.	35
Table 3	Tableau des perceptions brutes et des actions.	48
Table 4	Tableau résumé des variables utilisées dans le modèle.	55
Table 5	Comparaison des datasets CAV3D et LIA Pepper.	62
Table 6	Tableau résumé des paramètres pour CAV3D et le corpus Pepper.	62
Table 7	Tableau des performances (%) de notre modèle pour différents ensembles d'expériences, A_t est défini avec l'oracle Voice Activity Detection (VAD) dérivé des étiquettes de diarisation, ou avec la présence ou l'absence de Sound Source Localization (SSL).	63
Table 8	Tableau de la répartition des classes par niveau d'expression.	66
Table 9	Tableau des résultats des différentes approches pour $\tau = 5s$, timestep = 0.5s, $\eta = 2s$.	68
Table 10	Tableau de la répartition des valeur de l'IAB dans le jeu de données.	75

Table 11	Tableau des résultats des scores F1 et Area Under the Receiver Operating Characteristic Curve (AUROC) sur différentes combinaisons dans la fenêtre de temps optimale.	77
Table 12	Tableau récapitulatif des perceptions extraites.	79
Table 13	Questionnaire d'expérience utilisateur UMUX.	92

ACRONYMES

AES	Acceptability E-Scale
ANR	Agence Nationale de Recherche
ASR	Automatic Speech Recognition
AUROC	Area Under the Receiver Operating Characteristic Curve
BDI	Belief Desire Intention
CAV _{3D}	Co-located Audio-Visual streams with 3D tracks
DER	Diarization Error Rate
EM	Expectation Maximization
EST	Environment State Tracker
FA	False Alarm
FN	False Negative
FP	False Positive
GRU	Gate Recurrent Unit
HBU	Human Behavior Understanding
HRI	Human Robot Interaction
IAB	Interaction Acceptance Belief
LLM	Large Language Model
MS	Missing Detection
RNN	Recurrent Neural Network
ROS	Robot Operating System
SED	Sign of Engagement Decrease
SLAM	Simultaneous Localization and Mapping
SOT	Single Object Tracking
SOTA	State Of The Art
SPKE	Speaker Error
SSL	Sound Source Localization
SUS	System Usability Scale
TP	True Positive
UMUX	Usability Metric for User Experience
VAD	Voice Activity Detection

WER Word Error Rate

NOTATIONS

Type de variable	Notation	Exemple
Matrice	Lettre soulignée	<u>A</u>
Vecteur	Italique gras	v
Scalaire	Ni soulignée ni en gras	x
Variable aléatoire	Lettre majuscule	X
Réalisation d'une variable aléatoire	Lettre minuscule	x

Première partie

Introduction

INTRODUCTION

1.1 MOTIVATION

Au cours de la dernière décennie, nous avons été témoins d'une transition significative des robots, qui ont progressivement cessé d'être des entités confinées aux environnements industriels (ou de recherche) pour commencer à investir des espaces peuplés d'humains. Cette évolution va radicalement transformer notre interaction quotidienne avec la technologie. Par exemple, les robots aspirateurs, initialement perçus comme des gadgets futuristes, sont devenus des compagnons domestiques indispensables pour les plus chanceux, allégeant ainsi le fardeau des tâches ménagères. De plus, les avancées en intelligence artificielle ont permis la création d'agents conversationnels sophistiqués, intégrés dans des objets innovants qui semblent tout droit sortis des représentations du futur des décennies précédentes, à l'instar d'un Jibo (RANE, MHATRE et KURUP 2014). Ces dispositifs ne se contentent pas de répondre à nos commandes, mais engagent des interactions quasi naturelles, redéfinissant ainsi notre rapport aux machines. C'est dans ce contexte de transformation que le concept de robot social a vu le jour.

Au regard de ce paradigme, les robots sociaux (DAUTENHAHN 1998) représentent une nouvelle frontière dans la recherche. Ces entités automatisées, capables d'interactions autonomes avec l'homme dans un contexte social, dessinent de nouvelles opportunités pour optimiser la qualité de vie des utilisateurs. Grâce à l'intelligence artificielle, les échanges entre l'homme et le robot peuvent être enrichis, permettant ainsi au robot de cerner et de répondre aux besoins émotionnels et physiques de l'utilisateur. Ceci esquisse les contours d'une cohabitation plus personnalisée et efficace, où la technologie robotique s'intègre de façon harmonieuse dans la routine quotidienne, fournissant un accompagnement constant et adapté de l'individu. Ces robots sont amenés non seulement à naviguer de manière sécurisée parmi les objets et les individus, à saisir et manipuler des objets - autrement dit à interagir physiquement - mais ils doivent également être en mesure de communiquer avec les individus de la façon la plus naturelle possible, c'est-à-dire à interagir de façon cognitive. Parmi ces robots, les robots assistants socialement capables jouent un rôle prépondérant dans le domaine de la santé et du bien-être psychologique des personnes âgées (GONZÁLEZ-JIMÉNEZ, GALINDO et RUIZ-SARMIENTO 2012; PIASEK et WIECZOROWSKA-TOBIS 2018; NOGUCHI, KAMIDE et FUMIHIDE TANAKA 2020).

L'une des principales caractéristiques du robot social de nos jours est sa capacité à répondre aux besoins de l'utilisateur lorsque celui-ci lui en intime l'ordre. L'agent robotique analyse les demandes explicites de l'utilisateur telles qu'une demande d'assistance pour commander un plat dans un restaurant ou encore une question sur la météo de la semaine, et agit en conséquence. On appelle cela un comportement réactif, l'agent réagit aux événements induit par l'utilisateur. Ce type de comportement est efficace pour agir

sur des tâches d'assistance telles que les tâches de service à la clientèle. Par exemple, le robot de SHIOMI, SAKAMOTO et al. (2011) est conçu pour guider les visiteurs dans une station de train, offrant des informations claires et précises pour faciliter leurs déplacements. De même, KANDA, SHIOMI et al. (2009) ont développé un robot destiné à fournir des indications d'itinéraire et d'autres informations pertinentes dans un centre commercial, améliorant ainsi l'expérience des visiteurs en leur offrant une assistance personnalisée et immédiate. Plus récemment, C. S. CHEN, C. J. LIN et LAI (2022) ont introduit un robot pour le service alimentaire dans les restaurants de restauration rapide, optimisant les processus de commande et de livraison des repas pour une efficacité accrue. Cette réactivité confère à ces agents des capacités d'adaptation à de nombreuses situations variées. Toutefois, elle présente des limites notables. Par exemple, si un humain est exposé à un danger imminent mais ne sollicite pas explicitement l'aide du robot, le comportement réactif de l'agent se révélera insuffisant pour anticiper un scénario où l'individu pourrait se blesser. En outre, ce manque de prévoyance peut rendre les interactions avec le robot moins naturelles et plus mécaniques, en raison de son incapacité à anticiper l'évolution de son environnement.

Ainsi, chez l'humain, les comportements réactifs sont complétés par une autre famille de comportement, les comportements proactifs. La définition française d'un individu proactif est "celui qui anticipe les réponses à donner à un problème". Dans le domaine de l'interaction humain-robot, elle est définie comme *la capacité d'initier de manière autonome des actions anticipatoires, fondées sur le raisonnement, visant à influencer les individus et/ou leur environnement* (GROSINGER 2022), renforçant ainsi leur utilité et leur rôle social dans diverses situations. Par exemple, un robot social pourrait proposer des activités adaptées à un individu âgé pour stimuler son bien-être cognitif et physique, ou détecter et réagir de manière appropriée aux signes de détresse émotionnelle. En outre, l'adaptabilité des robots sociaux est cruciale. Ils doivent être capables de s'adapter aux préférences individuelles, aux habitudes et aux comportements spécifiques de chaque utilisateur. Cela implique non seulement une personnalisation dans leurs interactions, mais également une capacité à apprendre et à évoluer au fil du temps grâce à des mécanismes d'apprentissage. Ces robots peuvent fournir un soutien plus personnalisé et efficace, établissant ainsi une relation plus significative et durable avec les utilisateurs. Cette constante adaptation réside notamment dans l'analyse et le traitement des informations reçues par les récepteurs sensoriels du robot. Ces perceptions, semblables à l'humain, telles que la vision ou l'ouïe sont combinées, améliorées et utilisées pour adapter le comportement de l'agent au monde qui l'entoure.

Illustrons cela avec l'exemple d'une salle d'attente dans une unité hospitalière, qui, aux heures de pointe, est fréquentée par de nombreux patients, accompagnants et personnels hospitaliers. Un robot serait d'une grande utilité dans cet environnement, avec pour objectifs : prendre en compte les patients en attente depuis un certain temps et susceptibles d'éprouver de l'anxiété, engager la conversation pour atténuer leur stress, les informer des retards, les divertir pendant l'attente, et éventuellement les guider vers leur prochain rendez-vous médical. Pour mener à bien ces missions, le robot devrait être capable de voir, d'entendre et de communiquer sur diverses distances, de se déplacer de manière

sécurisée dans la salle et l'hôpital, d'approcher une personne ou un groupe de personnes, et de s'engager dans un dialogue multimodal face à face avec un patient, un membre de la famille, un membre du personnel, ou un groupe d'entre eux. Bien sûr, la conception de ces robots doit également répondre aux défis éthiques et sociaux importants. Il est crucial de garantir la confidentialité et la sécurité des données personnelles, tout en s'assurant que l'intégration de ces technologies respecte les normes sociales et culturelles établies.

Cette dissertation se penche sur le développement d'un tel robot social, spécialement conçu pour des interactions proactives avec les utilisateurs. Elle décrit les étapes nécessaires à sa réalisation depuis l'accumulation de perceptions du robot par ses capteurs, jusqu'à l'envoi d'ordres d'action à ses actionneurs. L'expérience se situe dans un contexte particulier : une salle d'attente d'hôpital où le robot est programmé pour une interaction dynamique avec les patients. Nous développerons un ensemble de stratégies exhaustives pour doter le robot Pepper (anciennement de Softbank Robotics, maintenant de United Robotics Group) d'un comportement socialement proactif. L'architecture que nous proposons pour cet objectif sera validée à travers une série d'expériences avec des humains en temps réel simulant le contexte hospitalier ainsi qu'une expérience dans un centre hospitalier. Le tout est développé en suivant les contraintes de latences inhérentes au système d'interaction humain-robot.

1.2 DÉFINITION DES CONCEPTS

Dans cette section, nous définissons les concepts principaux qui seront abordés tout au long de cette thèse. Certains de ces termes sont utilisés dans différents sens dans le domaine de l'interaction homme-robot. Certains termes plus spécifiques tels que la proactivité seront définis plus tard dans le manuscrit.

Terme	Définition
L'agent	représente l'individu, humain ou robot, capable d'effectuer une action.
L'utilisateur	est un individu qui interagit avec ou utilise l'outil.
Le participant	est un individu humain qui participe à l'expérience.
La tâche	est utilisée pour définir l'interaction globale, y compris le jeu ou l'événement auquel l'utilisateur est confronté avec le robot.
Le corpus	désigne un ensemble structuré de données textuelles ou multimédia qui sont utilisées pour développer, tester et évaluer des algorithmes et des modèles.
Interaction sociale	relation interpersonnelle entre deux individus (au moins), par laquelle les comportements de ces individus s'influencent mutuellement et se modifient chacun en conséquence.
Perceptions brutes	perceptions directement issues des capteurs du robot.
Perceptions raffinées	données brutes traitées et améliorées pour fournir des informations environnementales plus précises.

TABLE 1 – Tableau récapitulatif des définitions des termes utilisés dans l'étude

1.3 OBJECTIFS, LE CONTEXTE DU PROJET ANR μ DIALBOT

Cette thèse s'inscrit dans le cadre du projet de l'Agence Nationale de la Recherche μ DialBot, qui ambitionne d'intégrer activement des indices comportementaux humains dans la communication orale entre l'homme et le robot. Le projet μ DialBot vise à exploiter pleinement les riches informations fournies par les données audio et visuelles émises par les humains lors de leurs interactions avec les robots. En particulier, l'extraction de caractéristiques perceptives verbales et non verbales hautement informatives doit permettre d'améliorer la capacité de prise de décision des robots, leur permettant ainsi de prendre des tours de parole de manière plus naturelle et d'alterner entre des interactions en groupe ou à plusieurs et des dialogues en face à face selon les besoins.

L'intérêt pour les robots compagnons, capables d'assister les personnes dans leur vie quotidienne et de communiquer avec elles, s'étant récemment accru, ces robots sont désormais perçus comme de véritables entités sociales. Et leur utilité pour la santé et le bien-être psychologique des personnes âgées a été reconnue par plusieurs études récentes. Les patients, leurs familles et les professionnels de la santé apprécient le potentiel des robots, à condition que plusieurs barrières technologiques soient surmontées dans un avenir proche, notamment la capacité de se déplacer, de voir et d'entendre afin de communiquer naturellement avec les gens, bien au-delà de simples écrans tactiles et des commandes vocales.

Les résultats scientifiques et technologiques obtenus dans le cadre de ce projet sont implémentés sur un robot social. Ils ont été testés et validés dans plusieurs cas d'utilisation

au sein d'une unité hospitalière de jour. La collecte de données qui complète les tests in situ, fournit ainsi une base solide pour alimenter davantage de recherches et affiner les capacités des robots sociaux à répondre aux besoins des utilisateurs dans des environnements réels.

Nous nous sommes attachés à développer des scénarios permettant de gérer des interactions proactives fondées sur la perception, où un robot autonome fait face à des contextes impliquant plusieurs individus. Le robot devra choisir entre établir des interactions de groupe par le dialogue ou engager des échanges en tête-à-tête, et passer de manière proactive d'un mode à l'autre grâce à une perception multimodale fiable. Ces tâches nécessitent l'extraction efficace de caractéristiques audio et visuelles (telles que la clarté de la parole, la prosodie, le mouvement des yeux et de la tête, les expressions faciales, les mouvements des lèvres et de la tête, et les gestes manuels) à partir des données brutes des capteurs.

Cependant, il est difficile d'extraire de manière fiable ces caractéristiques avec un robot statique, écoutant et parlant à distance (3-4 mètres), en raison de diverses contraintes telles que le bruit acoustique, la réverbération, la superposition de sources audio, un éclairage insuffisant, une résolution d'image limitée, un champ de vision réduit de la caméra, des angles de vue non frontaux, des occlusions visuelles, etc. Par conséquent, la perception et la communication audiovisuelles doivent être activement menées. Face à une tâche spécifique telle qu'un dialogue en tête-à-tête, le robot doit être capable de détecter des événements sociaux, même éloignés, puis de s'approcher d'une personne ou d'un groupe pour assurer la collecte de données de qualité (par exemple, des vidéos frontales de visages et des signaux audio avec un bon rapport signal/bruit). Cela permet au robot de répondre et de réagir de manière adaptée aux interactions verbales et non verbales humaines, tout en respectant les normes de distanciation sociale. La photo 1 illustre l'environnement utilisé pour conduire les expériences.

Ainsi, μ DialBot va contribuer à la généralisation des compétences des robots vers des interactions homme-robot réalistes et non contraintes.

Plus précisément, l'objectif de μ DialBot est double : (i) le développement de méthodologies d'apprentissage statistique et profond pour la fusion de données, l'intégration sensorimotrice, la compréhension du comportement humain et le dialogue multimodal et (ii) la mise en œuvre de scénarios réalistes menant à une future génération de robots sociaux capables d'utiliser leurs capteurs embarqués (caméras et microphones) et actionneurs, pour percevoir les gens, analyser leurs comportements individuels et de groupe, se déplacer en toute sécurité vers un groupe ciblé de personnes, manifester de l'empathie et s'engager dans une communication en face à face.

Les algorithmes proposés doivent être implémentés sur un robot de nouvelle génération, et validés expérimentalement dans une des salles d'attente de l'hôpital Broca, peuplée de patients, d'accompagnateurs et de personnel médical. La tâche du robot sera d'engager le dialogue pour aider les patients et assister les soignants. Plusieurs tâches orientées vers des objectifs ont été identifiées, notamment : le rappel de rendez-vous, la discussion en cas de retard, les tests pré-examen, les visites guidées, les jeux, la sélection



FIGURE 1 – Image de la salle d'expérience de l'hôpital Broca.

de musique et la détection de comportements anormaux. Un défi majeur est d'effectuer une analyse de scène audio-visuelle, par exemple jusqu'à 4 mètres, pour comprendre globalement une situation sociale, puis agir de manière adéquate et opportune pour sélectionner un groupe de personnes, se déplacer vers elles et s'engager dans un dialogue de groupe. Une attention particulière sera portée aux questions éthiques, de sécurité et de confidentialité associées à la collecte, l'annotation, la diffusion et l'utilisation des données. Malgré les difficultés et la lourdeur supplémentaire afférente, le consortium a pu mettre en œuvre une collecte de données strictement conforme aux réglementations en vigueur concernant le stockage des données ainsi que les règles de consentement pour l'accès, l'utilisation et la diffusion des données.

Ainsi ce projet de thèse vise à étudier les interactions humain-robot en développant une technologie capable de comprendre et de participer activement à des situations sociales proactives. En respectant les aspects éthiques, de sécurité et de confidentialité, ce travail contribuera à explorer les capacités des robots sociaux dans un cadre de salle d'attente d'une unité hospitalière. Aussi les objectifs de cette thèse peuvent être résumés en fonction du projet μ DialBot dans lequel elle s'inscrit de la manière suivante :

- **O1. Développer un système d'interaction multimodal proactif dans le cadre d'une interaction humain-robot.**
- **O2. Tester et évaluer cette architecture dans un scénario de salle d'attente d'hôpital.**

1.4 CONTRIBUTIONS

Cette recherche a conduit à la conception d'un système multimodal d'interaction humain-robot dans le cadre d'une salle d'attente hospitalière capable d'interagir proactivement avec les patients, utilisant des systèmes robustes de suivi du comportement humain. Plus précisément :

- un modèle de différenciation temporelle du locuteur basé sur des concepts mathématiques et algorithmiques de principe couplés à deux types de perception, la localisation active de source audio (SSL, Sound Source Localization) et l'image plane capté par les caméras, avec une capacité d'adaptation à différentes conditions acoustiques sans phase d'entraînement ;
- le concept de 'croyance d'acceptation de l'interaction' (IAB, Interaction Acceptance Belief) dans le domaine de l'interaction homme-robot avec une approche de modélisation pour la prédiction de l'IAB et validée de manière approfondie en utilisant un ensemble de données réelles comprenant divers scénarios ;
- une architecture d'interaction humain-robot multimodale proactive, FlowAct, qui fonctionne comme une boucle sans fin asynchrone des capteurs du robot vers les actionneurs et organisée par deux contrôleurs : le suivi de l'état de l'environnement et le planificateur d'actions ;
- une implémentation de l'ensemble du système dans le robot social multimodal Pepper et la validation de son utilisation et logique à travers des expériences "d'exploration" dans un cadre de salle d'attente d'un complexe hospitalier.

1.5 ORGANISATION DES SECTIONS

chapitre 2 : Travaux connexes - Le deuxième chapitre se concentrera sur les travaux connexes à cette dissertation. Il commence par une introduction au domaine de l'interaction Homme-Robot HRI, puis définit le concept actuel de HRI, décrit les robots sociaux et leurs applications, les modèles d'HRI, le système d'exploitation pour robots (ROS) et le robot multimodal Pepper. La section suivante se concentre sur l'apprentissage machine, elle couvre à la fois les approches non neuronales comme les mixtures gaussiennes et les approches neuronales d'apprentissage profond. Ensuite, elle aborde les signaux de l'environnement, en particulier la reconnaissance automatique de la parole, la détection

et le suivi des personnes par vision, la diarisation du locuteur, et les signaux sociaux. La proactivité est également discutée, surtout dans la vision théorique de la psychologie organisationnelle, puis pragmatiquement en [HRI](#), et sa modélisation. Enfin, le contexte du projet μ DialBot est rappelé, pour montrer à quel niveau du projet cette thèse s'intègre.

chapitre 3 : Contributions - Le chapitre des contributions détaille les avancées et innovations apportées par ce travail. Il décrit le flux de perception, incluant l'extraction des perceptions brutes et un résumé des méthodes utilisées. Ensuite, le raffinement des perceptions est abordé avec des sous-sections consacrées à la parole, la vision, le développement d'un module de diarisation multimodale du locuteur dans le cadre de l'[HRI](#), l'analyse de l'engagement en [HRI](#), et l'analyse de l'acceptation de l'engagement. La section suivante présente FlowAct, un modèle d'interaction humain-robot, en expliquant son fonctionnement général, la perception, la modélisation de l'environnement, la prise de décision et la planification d'action, suivi d'un résumé. Enfin, la section des expérimentations décrit les expériences réalisées, notamment en laboratoire contrôlé et à l'hôpital Broca, pour valider les contributions théoriques et pratiques du travail.

chapitre 4 : Conclusion - Le chapitre final du document présente une synthèse des travaux réalisés et des perspectives futures. Elle commence par une conclusion générale, récapitulant les principaux résultats obtenus. Ensuite, elle énumère les contributions spécifiques de la recherche, détaillant les avancées apportées. La discussion qui suit offre une réflexion critique sur les résultats, mettant en lumière les points forts et les implications. La section aborde également les limitations rencontrées au cours de l'étude, en identifiant les aspects qui pourraient (devraient) être améliorés ou qui nécessitent des recherches supplémentaires. Enfin, les perspectives sont discutées, proposant des directions pour les travaux futurs et suggérant des pistes pour approfondir et étendre les recherches dans ce domaine.

1.6 PUBLICATIONS

Le contenu de cette thèse repose en partie sur les travaux publiés dans les articles suivants. Ils contiennent des résultats et des analyses qui ont été cruciales pour faire avancer les concepts et les expériences présentées dans cette thèse. Vous retrouverez ces articles intégralement discutés et analysés dans les sections appropriées de cette thèse :

- **Audio-Visual Speaker Diarization in the Framework of Multi-User Human-Robot Interaction**, Timothée Dhaussy, Bassam Jabaian, Fabrice Lefèvre, Radu Horaud ICASSP 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP);
- **Interaction acceptance modelling and estimation for a proactive engagement in the context of human-robot interactions**, Timothée Dhaussy, Bassam Jabaian, Fabrice Lefèvre, 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW);
- **FlowAct : Un système d'interaction homme-robot multimodal pro-actif avec un flux continu de perception et des sous-systèmes d'action modulaires**, Timothée

- Dhaussy, Bassam Jabaian, Fabrice Lefèvre, 2024 *Affects, Compagnons Artificiels et Interactions Workshop (WACAI)*;
- **FlowAct :A Proactive Multimodal Human-robot Interaction System with Continuous Flow of Perception and Modular Action Sub-systems**, Timothée Dhaussy, Bassam Jabaian, Fabrice Lefèvre, 2024 *ArXiv.org*;

Deuxième partie

Travaux connexes

TRAVAUX CONNEXES : INTERACTION HUMAIN-ROBOT ET PROACTIVITÉ

2.1 INTRODUCTION

Ce chapitre propose une synthèse de la littérature en [HRI](#), explorant le cadre théorique et l'état actuel des recherches dans les divers domaines concernés par nos objectifs, afin de définir le contexte spécifique du travail de thèse présenté. Il est divisé en cinq sections, chacune constituant un élément essentiel pour la création d'un agent interactif proactif.

Dans la première section, nous aborderons le concept de l'interaction humain-robot, en détaillant sa définition, son importance dans le contexte social à travers les robots à capacités sociales, le robot spécifique utilisé pour cette étude, ainsi que le système d'exploitation logiciel [ROS](#) qui va accueillir l'architecture proposée pour nos travaux.

La deuxième section abordera la littérature relative à l'apprentissage machine à travers les approches non neuronales et par réseaux de neurones.

La troisième section est dédiée à l'élaboration des signaux audiovisuels permettant de réaliser une représentation la plus exacte possible de la scène pour l'agent, ainsi qu'aux outils nécessaires pour ce faire.

La quatrième section se concentre sur la notion de proactivité, en explorant sa signification tant en psychologie qu'en informatique. Nous examinerons diverses approches de modélisation de la proactivité, ainsi que des concepts liés à l'initiation proactive des interactions.

Enfin, la cinquième section présentera plus en détail le projet μ Dialbot dans lequel s'inscrit cette thèse.

2.2 L'INTERACTION HUMAIN-ROBOT

2.2.1 Définition de la [HRI](#)

L'interaction humain-robot ([HRI](#)) constitue un domaine de recherche interdisciplinaire en rapide évolution qui se situe à l'intersection de plusieurs disciplines, notamment l'informatique, la robotique, les sciences cognitives, la psychologie sociale, et même l'éthique (FONG, NOURBAKHSI et DAUTENHAHN 2003; DAUTENHAHN 2007). Ce champ d'étude multidisciplinaire aspire à concevoir des systèmes robotiques sophistiqués capables de comprendre, d'interagir et de collaborer avec les êtres humains de manière efficace et significative. Cette aspiration est catalysée par le potentiel transformationnel que les robots sociaux présentent dans divers secteurs applicatifs, allant de la santé et de l'éducation à des domaines plus spécialisés tels que la recherche en environnements hostiles. Cette multidisciplinarité caractéristique de la [HRI](#) s'exprime à travers différents domaines tels que l'informatique, l'ingénierie et les sciences sociales, comme illustré par la figure 2. Cette interdisciplinarité rend la maîtrise de la [HRI](#) particulièrement complexe, nécessitant

une compréhension approfondie et intégrée de multiples disciplines pour développer des systèmes efficaces et intuitifs pour leurs utilisateurs. Ce caractère multidisciplinaire exige des chercheurs et des praticiens une approche holistique et collaborative pour aborder les nombreux défis du domaine.

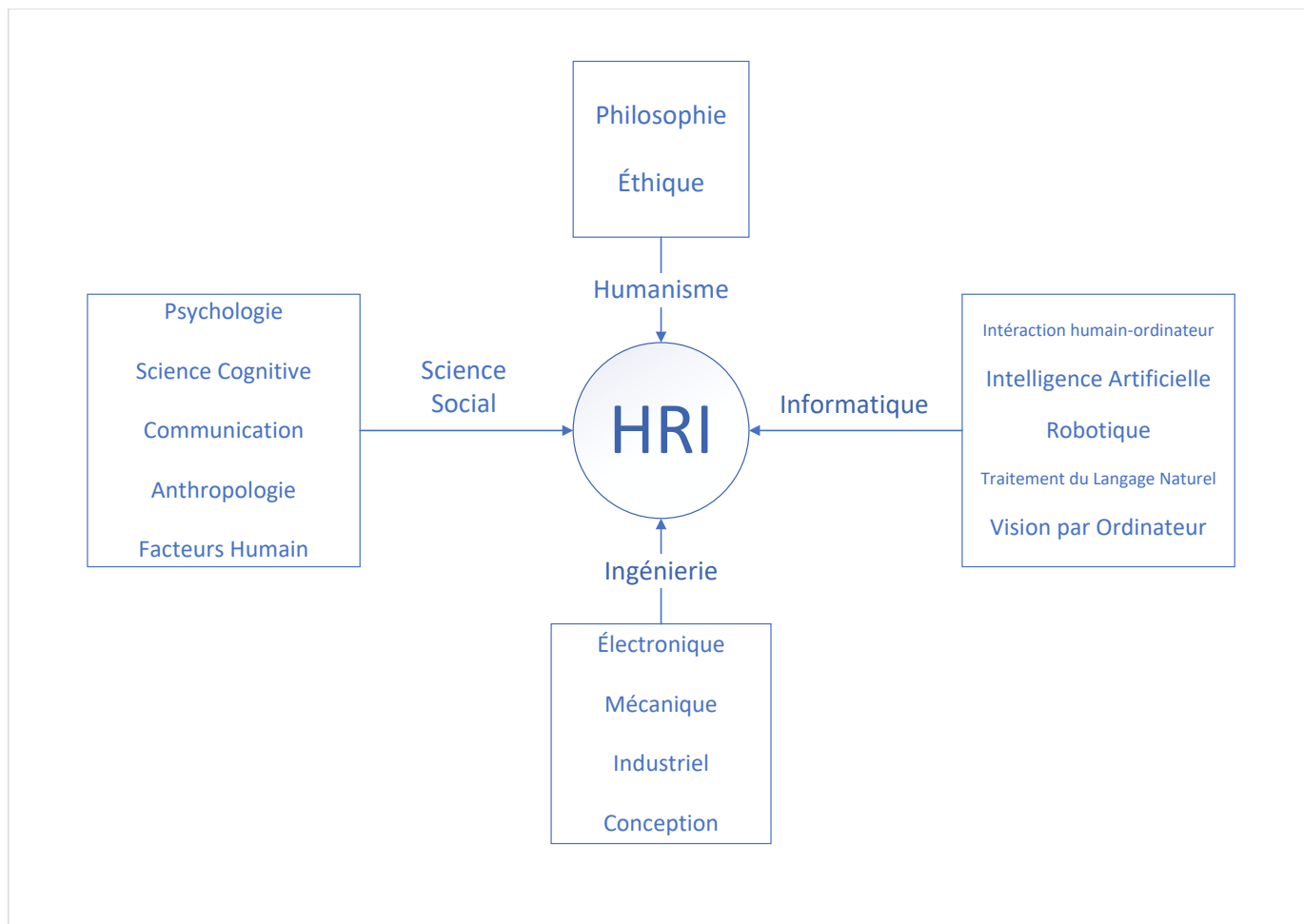


FIGURE 2 – Diagramme des champs des disciplines impliquées dans l'Interaction Humain-Robot (GOLAM RASHED 2016).

2.2.2 Le robot social

2.2.2.1 Définition

Bien que l'émergence des robots sociaux soit un phénomène relativement récent, l'idée d'attribuer des capacités sociales à des objets de notre environnement remonte à bien plus longtemps. Cette tendance à attribuer des caractéristiques humaines ou émotionnelles à des objets inanimés est connue sous le nom d'anthropomorphisme (URQUIZA-HAAS et KOTRSCHAL 2015). Divers exemples historiques illustrent cette tendance. On peut citer, par exemple, le héron d'Alexandrie (DAUTENHAHN 2007), conçu en 62 après J.-C. en Égypte, ou encore l'horloge en forme d'éléphant d'Ismail al-Jazari (ELICES 2020),

datant du 12^{ème} siècle en Turquie, avec ses figures humaines, et les Poupées de Karakuri (GRAEFE et BISCHOFF 2009) du 17^{ème} siècle au Japon, conçues pour servir le thé.

Avec les progrès actuels en matière de matériel et de logiciel informatique, l'humanité poursuit la création de telles machines anthropomorphes, tant pour des applications fonctionnelles que pour le divertissement. La robotique moderne ne se concentre plus uniquement sur l'aspect mécanique des machines, mais accorde également une importance particulière à l'esprit ou à l'intelligence du robot. De nos jours, la robotique est un domaine en constante expansion. En synergie avec l'Interaction humain-ordinateur et la psychologie, elle a contribué à l'émergence du domaine de l'Interaction humain-robot (HRI), un champ de recherche relativement nouveau et, pourrait-on dire, encore en phase de développement. Dans ce contexte nouveau, le terme de robot social a commencé à être largement utilisé.

Il existe plusieurs définitions utilisées dans la littérature scientifique pour qualifier un robot social :

- *Socialement évocateur* : Robots qui s'appuient sur la tendance humaine à l'anthropomorphisation et qui capitalisent sur les sentiments évoqués lorsque les humains nourrissent, soignent ou impliquent leur "création" (BREAZEAL 2003; BREAZEAL 2004).
- *Situés socialement* : Robots entourés d'un environnement social qu'ils perçoivent et auquel ils réagissent. Les robots socialement situés sont capables de faire la distinction entre d'autres agents sociaux et divers objets dans l'environnement (FONG, NOURBAKHSI et DAUTENHAHN 2003).
- *Sociable* : Robots qui s'engagent de manière proactive avec les humains afin de satisfaire des objectifs sociaux internes (pulsions, émotions, etc.). Ces robots nécessitent des modèles profonds de cognition sociale (BREAZEAL 2003; BREAZEAL 2004).
- *Socialement intelligents* : Robots présentant des aspects de l'intelligence sociale humaine, basés sur des modèles éventuellement profonds de la cognition et de la compétence sociale de l'homme (DAUTENHAHN 1998).
- *Robots socialement interactifs* : Robots pour lesquels l'interaction sociale joue un rôle clé dans l'HRI de pair à pair, différents des autres robots qui impliquent une HRI "conventionnelle", tels que ceux utilisés dans les scénarios de téléopération. (FONG, NOURBAKHSI et DAUTENHAHN 2003)

Ces définitions, bien que distinctes, se complètent et se lient de manière significative pour former une compréhension holistique des robots sociaux. Elles suggèrent qu'un robot social idéal ne se limite pas à une seule facette de l'intelligence ou de la capacité interactionnelle. Au contraire, il représente une convergence de plusieurs aspects émotionnels, cognitifs, perceptifs et interactifs – qui sont tous essentiels pour créer une expérience sociale HRI authentique. Il est intéressant de noter que la proactivité, comportement humain contraire à la réactivité, est introduite dans la définition d'un robot social. On retrouve ces définitions du robot social dans l'espace conceptuel des approches HRI de DAUTENHAHN (2007). La figure 3 présente les trois différentes perspectives sur la HRI discutées dans cette section. Elle met en lumière l'accent mis dans différentes approches utilisant différentes définitions du comportement social des robots et forme un espace conceptuel des approches HRI.

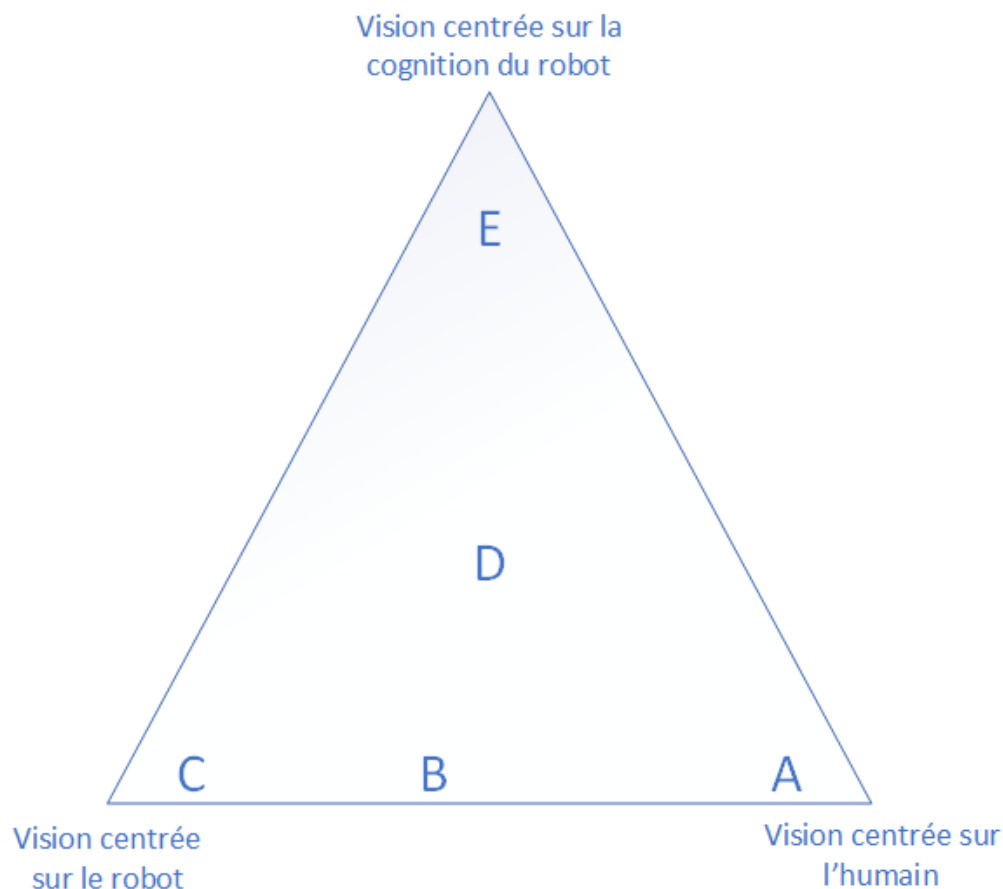


FIGURE 3 – Diagramme de l'espace conceptuel des approches HRI d'après (DAUTENHAHN 2007) A, socialement évocateur ; B, socialement situé ; C, sociable ; D, socialement intelligent ; E, socialement interactif.

2.2.2.2 Conception d'un robot social

Dans la conception du robot social, la forme et la fonction du robot sont étroitement liées et doivent être envisagées conjointement. L'anthropomorphisme en robotique, qui implique l'attribution de caractéristiques humaines aux robots, joue un rôle crucial dans la manière dont ces machines sont conçues et perçues. En dotant les robots d'aspects humains, tant dans leur apparence que dans leur comportement et leurs interactions, les chercheurs cherchent à les rendre plus familiers et acceptables pour les utilisateurs humains. Cette approche vise à faciliter l'intégration des robots dans des contextes sociaux et professionnels variés.

Cependant, cette tendance à humaniser les robots nous amène au concept de la vallée de l'étrange, introduit par MORI, MACDORMAN et KAGEKI (2012) (figure 4) en 1970. Selon cette théorie (non démontrée mais inspirante), lorsque la ressemblance d'un robot avec l'être humain augmente, la réaction émotionnelle des personnes passe de la positivité à la répulsion, si ses performances ne sont pas suffisamment convaincantes. Ce phénomène souligne un défi majeur dans la conception robotique : trouver le juste équilibre entre

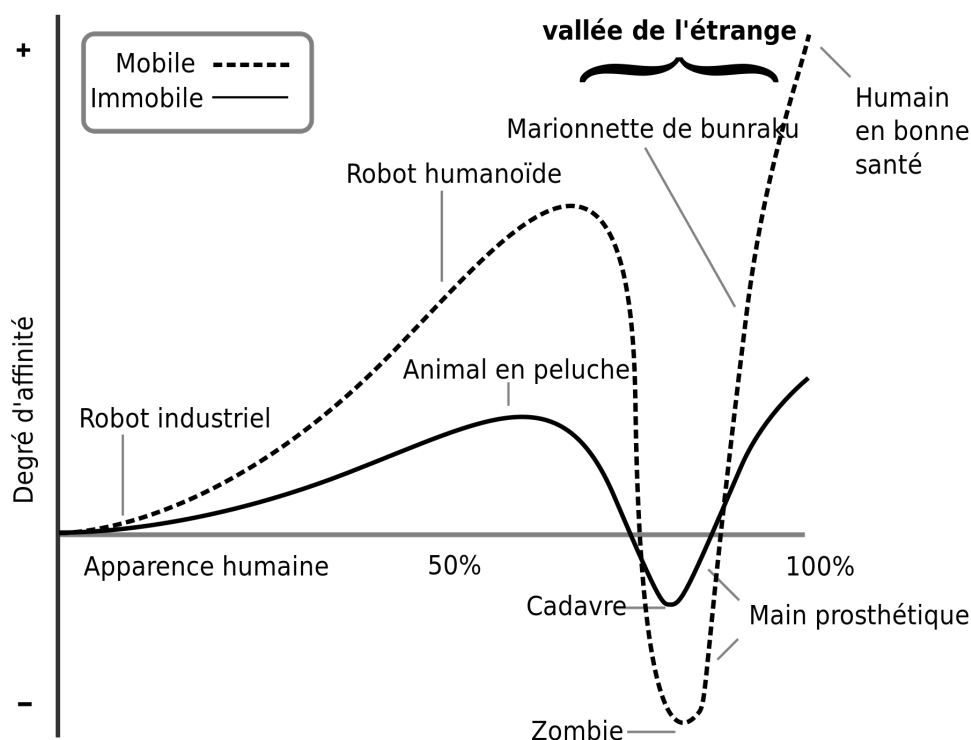


FIGURE 4 – Représentation de la vallée de l'étrange de MORI, MACDORMAN et KAGEKI (2012).

anthropomorphisme et acceptabilité, afin d'éviter les réactions négatives associées à une potentielle vallée de l'étrange.

Les chercheurs en HRI explorent une diversité de formes robotiques (figure 5). Cela inclut les androïdes et les humanoïdes, tels que NAO (JANISZEK et al. 2011), ARI (COOPER et al. 2020), Quori (SPECIAN et al. 2022), Moxie (HURST et al. 2020) et Miroki (TOOLS 2023), ainsi que les robots zoomorphes comme AIBO (KAPLAN 2000), RoboParrot (SHAYAN et al. 2016) ou encore RobotCat (« *Robotic cat is the perfect pet for seniors* » 2015). En outre, les conceptions minimalistes, telles que Jibo (RANE, MHATRE et KURUP 2014) et Tessa (*Tessa Robot - Zorgrobot voor zelfregie en zelfstandig wonen s. d.*), offrent également des perspectives uniques. Chaque catégorie présente des défis spécifiques en termes de capacités d'interaction et d'intégration dans des contextes variés. Récemment, l'intérêt s'est porté également sur les "robjects", des artefacts robotiques (Pranav DIXIT 2013) interactifs basés sur des *objets*, élargissant ainsi le champ du design possible du robot social.

2.2.3 Les applications en HRI

Les applications envisagées dans le cadre de la HRI englobent un large éventail de services généralement assurés par des êtres humains, avec des champs d'application variés. Un des principaux domaines concernés est celui des services, où l'on retrouve des robots dans des rôles tels que serveurs dans des restaurants ou guides dans des

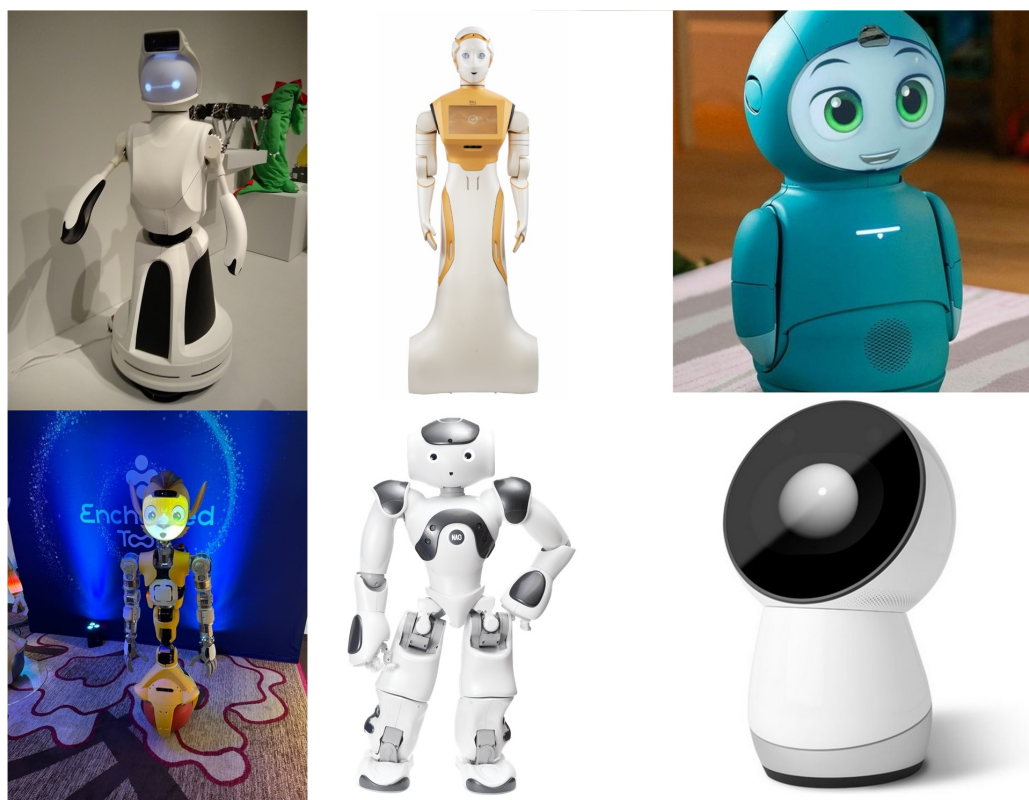


FIGURE 5 – Exemples de robots sociaux, Quori (SPECIAN et al. 2022), Ari (COOPER et al. 2020), Moxie (HURST et al. 2020), Miraki (TOOLS 2023), Nao (JANISZEK et al. 2011), Jibo (RANE, MHATRE et KURUP 2014).

centres commerciaux (KANDA, SHIOMI et al. 2009) et des gares (SHIOMI, SAKAMOTO et al. 2011).

Le secteur du divertissement est également un domaine en pleine expansion pour les robots sociaux (PASSLER BATES et YOUNG 2020; MAHDI et al. 2020). Dans ce contexte, leur rôle principal consiste à divertir l'utilisateur, soit en jouant avec lui, soit en engageant des conversations.

Dans le domaine de la santé, les robots jouent un rôle crucial, en particulier pour aider les personnes atteintes d'autisme (SHAYAN et al. 2016; HAN et al. 2018). Ils peuvent fournir un soutien personnalisé et contribuer à améliorer la qualité de vie des patients.

L'éducation représente une autre catégorie majeure dans laquelle les robots sont de plus en plus présents. Ils participent activement à la diffusion des connaissances et à l'assistance pédagogique (J. HAN et al. 2009; PARK et al. 2018; GUGGEMOS, SEUFERT et SONDEREGGER 2020). Ces robots peuvent intervenir en tant qu'assistants enseignants, facilitant l'apprentissage et rendant les cours plus interactifs et engageants pour les élèves.

Ainsi, les robots de la HRI trouvent leur utilité dans divers secteurs, transformant la manière dont les services sont rendus et enrichissant les interactions humaines par leur présence et leurs capacités technologiques avancées.

2.2.4 Les modèles HRI

La mise au point d'un modèle d'interaction humain-robot de haute performance implique l'intégration de modules capables de gérer simultanément et efficacement diverses tâches. Ces tâches comprennent la collecte et l'amélioration des données sensorielles, la modélisation de l'environnement en utilisant à la fois la mémoire et les informations acquises en temps réel, ainsi que la planification et l'exécution d'actions.

Dans PRESCOTT et al. (2019a) les auteurs étudient le concept de voyage mental dans le temps, qui se définit comme la capacité de projeter l'esprit dans le passé pour retrouver des souvenirs d'événements passés, et dans l'avenir pour imaginer des événements futurs possibles, en HRI et proposent une implémentation pour le robot iCub. Le modèle qu'ils proposent distingue deux niveaux de traitement de l'information : d'une part, la mémoire des capteurs, qui traite la perception sensorielle brute, et d'autre part, la mémoire associative, qui interprète des informations plus complexes à partir des données recueillies par les capteurs telles que la reconnaissance de personne. Dans leur article, MOULIN-FRIER et al. (2018) proposent une organisation d'architecture cognitive basée sur le Contrôle Adaptatif Distribué (Paul VERSCHURE, VOEGTLIN et DOUGLAS 2003; P.F. VERSCHURE, PENNARTZ et PEZZULO 2014), qui gère le traitement des états du monde ou extéroception, du soi ou interoception, et de l'action.

Le projet ADAPTA et son modèle RoboCog (ROMERO-GARCÉS et al. 2015) a mis en place un robot vendeur, capable de persuader des clients potentiels de s'approcher d'un stand de vente. Le modèle cognitif utilisé intègre une couche exécutive qui récupère les perceptions et transmet les commandes aux sous-modules d'actions, tels que les modules de navigation, d'expression, ou encore de conversation. Ce robot était capable d'identifier les clients, d'évaluer leur volonté de suivre, et de répondre à des questions spécifiques. L'architecture BRILLO (ROSSI et al. 2022), conçue pour un robot barman social, adopte une organisation en trois couches : la couche d'exécution, la couche de prise de conscience du contexte et de prise de décision, et la couche des perceptions, toutes implémentées avec ROS (cf section 2.2.5 ci-dessous), une collection efficace de bibliothèques logicielles et d'outils pour développer des applications robotiques.

L'architecture cognitive de LEMAIGNAN, WARNIER et al. (2017) est composée de différents modules, il interagit avec une couche sensorimotrice de bas niveau. On y retrouve notamment un module dédié au raisonnement géométrique et temporel, appelé Spark (Spatial Reasoning & Knowledge) (SISBOT, ROS et ALAMI 2011). Il fonctionne comme un évaluateur de situation qui génère des connaissances symboliques à partir de la géométrie de l'environnement en tenant compte des relations entre les objets, les robots et les humains. La gestion des connaissances repose sur un serveur central (LEMAIGNAN, ROS et al. 2010), qui stocke les connaissances au fur et à mesure qu'elles sont produites par chacun des autres composants délibératifs. Le principal atout de cette architecture réside dans sa capacité à représenter la scène. Toutefois, comme le soulignent les auteurs, elle peut s'avérer difficile à implémenter pour des cas d'utilisation de spectre large et est principalement adaptée aux tâches de manipulation interactive ainsi qu'aux activités qui y sont associées.

Les architectures précédemment décrites sont spécifiquement conçues et adaptées à une tâche d'application particulière, les rendant complexes et peu flexibles à implémenter.

ter pour notre cas d'utilisation. Cependant, elles partagent pour la plupart une structure similaire de traitement en trois couches : la couche perceptive, la couche de représentation et la couche d'action. Stratégie que nous avons conservée dans le développement de l'architecture FlowAct utilisée pour le projet μ Dialbot.

2.2.5 ROS

Le Système d'Exploitation de Robot (ROS¹, Robot Operating System), est un outil flexible conçu pour faciliter le développement de logiciels pour les robots. Il offre un ensemble de services standardisés qui sont couramment utilisés dans une variété de robots, permettant ainsi aux développeurs d'être plus efficaces sur les détails de bas niveau de la communication et du contrôle (QUIGLEY et al. 2009).

ROS est un outil essentiel pour la recherche et le développement dans le domaine de la robotique car il offre une architecture de communication inter-processus qui facilite l'échange d'informations entre les différents composants d'un robot. De plus, ROS fournit une grande variété de bibliothèques et d'outils pour aider les développeurs à construire et à déboguer leurs systèmes.

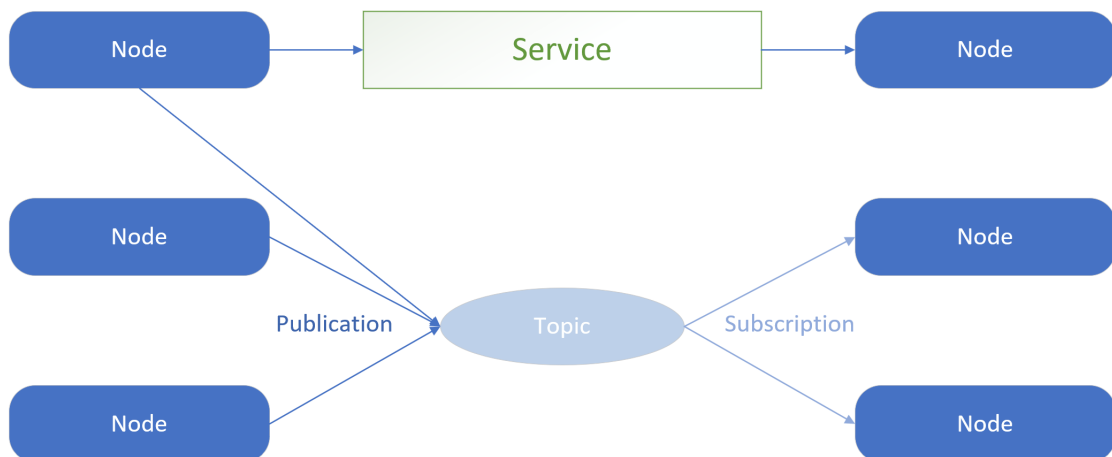


FIGURE 6 – Diagramme de l'architecture de la communication dans le système ROS.

ROS fonctionne comme un cadre de communication entre diverses parties d'un système robotique. Il est basé sur une architecture (figure 6) de type "publish-subscribe" et "client-server", ce qui signifie que les différents composants d'un robot, appelés "nœuds", peuvent publier des messages sur des "sujets" ou fournir des "services". Plus précisément :

- Les "nœuds" sont des processus exécutables qui communiquent les uns avec les autres via le noyau ROS. Chaque nœud est conçu pour effectuer une tâche spécifique, comme contrôler un moteur, lire les données d'un capteur, effectuer un calcul... Les nœuds peuvent communiquer entre eux en publiant des messages sur des sujets, en fournissant des services ou en utilisant des actions.

1. <https://www.ros.org/>

- Les "sujets" (ou topics) sont des canaux de communication nommés sur lesquels les nœuds peuvent publier des messages. Les autres nœuds qui sont intéressés par ces informations peuvent s'abonner à ces sujets pour recevoir les messages. Cela permet une communication asynchrone entre les nœuds.
- Les "services" sont un autre moyen de communication entre les nœuds qui permet une interaction synchrone. Un nœud fournit un service lorsqu'il peut effectuer une tâche spécifique avec une entrée et une sortie définies. D'autres nœuds peuvent appeler ce service et attendre qu'il se termine.

En outre, ROS fournit des outils pour aider à la visualisation, au débogage et à la gestion des nœuds et des communications. Par exemple, "rviz" est un outil de visualisation 3D qui peut afficher les données des capteurs et l'état du robot. "roscop" est un outil qui permet d'enregistrer et de rejouer les données des sujets, ce qui est très utile pour le débogage et le test. Il existe également Gazebo, un simulateur de robotique en 3D, utilisé avec ROS pour développer et tester des algorithmes de robotique. Il permet de créer des environnements réalistes et de modéliser des interactions physiques, facilitant le prototypage rapide et les itérations fréquentes sans besoin de matériel physique. L'intégration avec ROS permet une communication bidirectionnelle fluide, ce qui est essentiel pour le contrôle et l'observation des simulations. Gazebo supporte la simulation de plusieurs robots interagissant ensemble.

ROS utilise une architecture modulaire, tirant parti d'une structure basée sur des paquets, qui favorise le développement, le partage et la réutilisation du code dans divers projets de robotique. Cependant, ROS présente des limites. Sa dépendance envers les systèmes d'exploitation standards implique souvent la nécessité de recourir à des ordinateurs à usage général. Cela entraîne une augmentation des coûts et limite les capacités de calcul en temps réel. Bien que ROS 2 tente de pallier ces limitations, il n'y parvient pas totalement (MACENSKI et al. 2022). Les quelques paquets de ROS 1 compatibles avec Pepper nous ont fait choisir la première version plutôt que la deuxième. Lorsque nous parlerons de ROS dans la suite du manuscrit, nous ferons donc référence à ROS 1, sauf indication contraire.

Au fil du temps, ROS a été enrichi de paquets essentiels tels que MoveIt! (COLEMAN et al. 2014) un framework de planification de mouvements robotiques se concentrant principalement sur la planification et l'optimisation des mouvements en environnement complexe et ros_control (CHITTA et al. 2017) un ensemble de bibliothèques et d'outils pour le contrôle matériel et logiciel des robots, capable d'exécuter les trajectoires fournies par MoveIt, augmentant ainsi le potentiel de l'outil. Ce dernier a été très bien accueilli par la communauté scientifique et son utilisation s'étend à divers domaines de la robotique. On le retrouve notamment dans les robots sociaux (ROSSI et al. 2022), la sécurité de l'Internet des objets (ROY et al. 2022), la communication des drones (H. LEE et al. 2021), ainsi que dans la planification des tâches et des mouvements robotisés (RAJENDRAN, V et O'BRIEN 2022).

En somme, ROS fournit une structure et des outils pour faciliter le développement de systèmes robotiques complexes. Sa flexibilité et sa modularité permettent aux développeurs de construire des robots avec une variété de capacités et de comportements.

2.2.6 *Pepper un robot multimodal*

2.2.6.1 *Présentation générale*

Initié en juin 2014 par Aldeberan², Pepper se présente comme un robot humanoïde innovant et captivant, le premier de sa catégorie à être équipé de compétences émotionnelles. Conçu pour servir d'assistant, Pepper est en mesure de percevoir et de répondre aux émotions humaines fondamentales grâce à une interface sophistiquée et des fonctionnalités avancées pour interagir avec son environnement.

Aldeberan/Softbank Robotics s'est positionné depuis quelques années comme un leader international en robotique, rassemblant plus de 500 collaborateurs dans des bureaux situés à Paris, Tokyo, San Francisco, Boston et Shanghai. Plus de 20 000 robots SoftBank Robotics, y compris NAO, Pepper et Romeo, sont déployés dans plus de 70 pays, dans des secteurs aussi variés que la recherche, l'éducation, la vente au détail, le tourisme, la santé ou encore le divertissement. La stratégie d'URG quant au futur de Pepper suite au rachat n'est pas encore clairement perceptible par le grand public.

À l'heure actuelle, Pepper est accessible au grand public au Japon et aux entreprises en Europe. Il est déjà employé par de nombreuses sociétés (SoftBank, Nestlé, Renault, Carrefour, Costa, Uniqlo...) pour accueillir les clients, générer du trafic, fournir des informations sur les produits et les services ou recueillir des données. Ses nombreux capteurs (figure 7) lui permettent de comprendre son environnement et de s'adapter en conséquence. Pepper est capable de se déplacer, d'éviter les obstacles, d'identifier les sons, de vous suivre et de se recharger de manière autonome.

Le robot Pepper est doté de 20 degrés de liberté, qui lui confèrent des mouvements naturels et expressifs. Il est équipé d'une tablette pour ajouter des interactions homme-machine graphiques et tactiles, et dispose d'une synthèse et d'une reconnaissance vocale disponibles dans 12 langues. Il est également équipé de modules de perception pour reconnaître et suivre les personnes, de plusieurs zones tactiles, de LEDs, de capteurs sensoriels et de micros pour des interactions multimodales. De plus, il est équipé de capteurs infrarouges, de pare-chocs, d'une centrale inertielle, de caméras 2D et 3D et de sonars pour une navigation omnidirectionnelle et autonome. Il pèse 28 kg et a une autonomie d'environ 12 h en sortie d'usine. En termes de sécurité, Pepper est équipé d'un système de pare-chocs ainsi qu'un système d'anti-collision, désactivable, basé sur des lasers et sonars.

Pepper est livré avec les logiciels et les ressources nécessaires pour les chercheurs et les enseignants. Il fonctionne sous NAOqi OS, un système d'exploitation basé sur Unix. Le framework NAOqi fournit les bases de programmation pour le développement d'applications pour Pepper en Python et C++ principalement. Plusieurs SDK sont fournis pour prendre le contrôle de Pepper et développer des comportements et des applications.

2.2.6.2 *En recherche*

Le robot Pepper a été étudié dans divers contextes de recherche depuis son introduction. F. TANAKA et al. (2015) ont décrit le développement d'une application éducative

2. Aldebaran a été rachetée par SoftBank Robotics, puis par l'entreprise allemande United Robotics Group (URG) en 2022

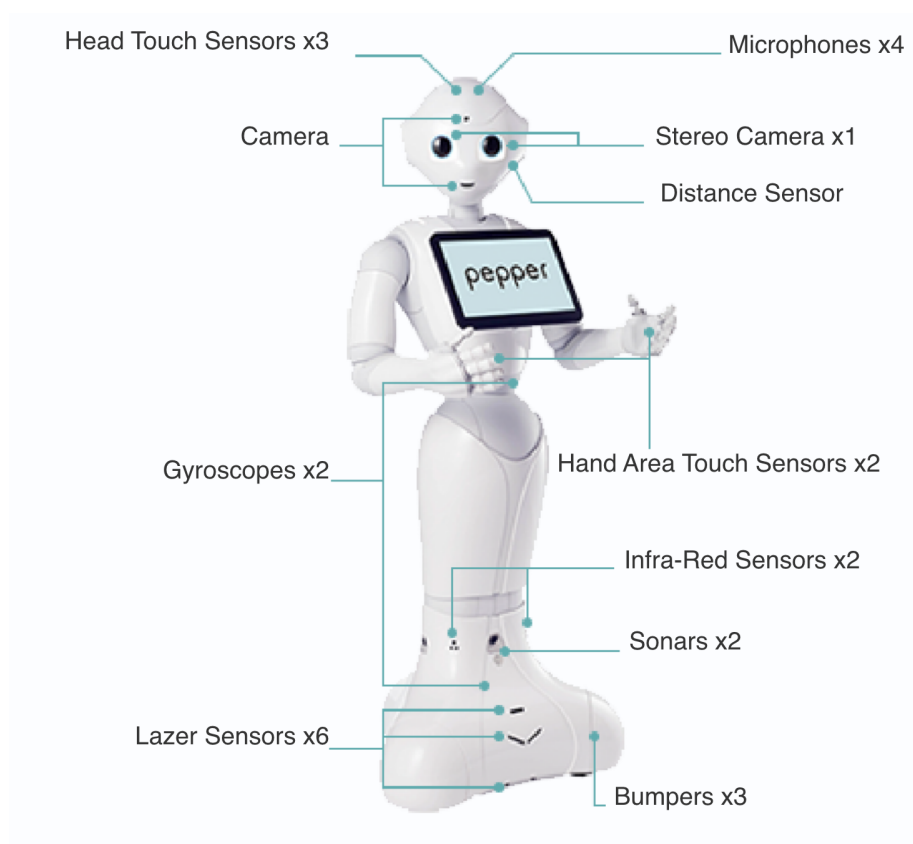


FIGURE 7 – Représentation des capteurs du robot Pepper.

pour le robot, conçu comme un agent éducatif qui apprend aux côtés des enfants. Ils envisagent une évolution future du contenu exploitant la capacité de Pepper à intégrer des extensions logicielles et du contenu stocké dans le cloud, anticipant une base IT large, diverse et adaptable. PERERA et al. (2017) ont été parmi les premiers à intégrer ROS sur Pepper, améliorant ainsi son autonomie grâce à l'usage de ROS et de services cloud. Cette intégration a significativement renforcé les capacités de navigation et de reconnaissance vocale de Pepper. Dans l'étude menée par GARDECKI et al. (2018), le robot a également été utilisé en tant que réceptionniste. Les auteurs de cette recherche ont opté pour l'intégration d'une caméra et d'un microphone externes afin d'optimiser les performances. On retrouve également une proposition d'architecture robotique MaestROB (MUNAWAR et al. 2018) implémentée à l'aide de ROS. Elle comprend plusieurs couches de composants de traitement, notamment la perception et le contrôle de l'actionnement au niveau inférieur. Aux niveaux supérieurs, elle intègre un planificateur symbolique et des API pour les fonctions cognitives et la compréhension sémantique. En 2020, GUGGEMOS, SEUFERT et SONDEREGGER (2020) ont analysé l'acceptation de Pepper dans un contexte d'enseignement supérieur, montrant que les étudiants évaluaient positivement la présence du robot en termes de confiance et d'adaptabilité, bien que leur intention de l'utiliser de manière autonome reste modérée, l'intention comportementale n'atteignant que 36,6% du maximum théorique. En 2021, FORBRIG, BUNDEA et BADER (2021) ont adapté Pepper

pour interagir avec des patients ayant subi un AVC, en développant des outils logiciels permettant une interaction plus flexible et ciblée.

Ces exemples montrent l'évolution de Pepper de simple outil éducatif à un acteur clé dans des scénarios interactifs complexes. Néanmoins, les capacités sensorielles de Pepper restent limitées pour des tâches telles que la localisation (PERERA et al. 2017). En 2024, on retrouve encore des études telles que Y. LI, HINDRIKS et KUNNEMAN (2024), s'attellant à des problèmes spécifiques de la partie audio de Pepper : élimination de l'écho, du chevauchement vocal ou encore du bruit parasite des ventilateurs. L'objectif est de filtrer ces bruits afin de laisser le micro continuellement ouvert, permettant un schéma de prise de parole plus naturel où l'humain peut interrompre le robot. Dans la section des limitations de sa thèse, BUYUKGOZ (2022) décrit les capacités de Pepper pour manipuler son environnement comme restreintes, en raison de son incapacité à saisir et déplacer des objets. Malgré la présence d'études sur l'utilisation de Pepper dans des tâches d'interactions avec les humains, on peut également noter la présence de recherches s'attachant à améliorer voire corriger des faiblesses de Pepper. Cela témoigne d'une certaine incompatibilité du robot pour des tâches complexes d'interaction humain-robot.

L'arrivée des modèles neuronaux à apprentissage profond vient renforcer les capacités des agents robotiques, X. CHEN et al. (2024) démontre que l'utilisation de Whisper d'OpenAI permet de réduire le taux d'erreur de mot (Word Error Rate (WER)) de 1.716% et le temps de calcul moyen de 2.639 secondes par rapport à Google Cloud Automatic Speech Recognition (ASR). L'utilisation de ChatGpt est également efficace pour répondre de manière contextuelle, détecter les commandes vocales et exécuter les instructions (BILLING, ROSÉN et LAMB 2023).

2.3 L'APPRENTISSAGE MACHINE

2.3.1 Introduction

L'intelligence artificielle, ainsi que son sous-domaine, l'apprentissage automatique, constituent des fondements essentiels pour le développement des robots sociaux et leur capacité à engager des interactions significatives avec les humains. Ces technologies confèrent aux robots la capacité de décoder et de traiter les comportements, les émotions et les intentions humaines, permettant ainsi des interactions plus fluides et adaptées. En particulier, l'apprentissage automatique dote les robots de meilleures capacités en leur permettant une meilleure perception de la scène, à travers la détection des utilisateurs, la reconnaissance des paroles prononcées et la détection des signaux sociaux.

Dans le cadre du projet μ DialBot, ces technologies permettent aux robots non seulement d'acquérir des compétences à partir de leurs interactions, mais aussi d'affiner leurs réponses, de s'adapter aux préférences des utilisateurs et de développer des modules de perception renforçant la représentation de la scène. Cela contribue également à améliorer la prise de décision en fonction de cette dernière. Ainsi, l'intégration de l'apprentissage automatique dans le projet μ DialBot s'impose comme une nécessité pour optimiser la qualité et la pertinence des interactions homme-machine.

2.3.2 Les approches non neuronales

Avant même que l'apprentissage profond ne révolutionne le paysage technologique, les techniques classiques d'apprentissage automatique s'étaient déjà fermement ancrées, et elles demeurent à ce jour des piliers incontournables dans le domaine de l'informatique. Parmi elles, l'apprentissage supervisé se distingue par ses méthodes de classification et de régression, incluant l'algorithme des plus proches voisins (DIXON 2006), la régression logistique (C. M. BISHOP 2007), la régression linéaire (C. M. BISHOP 2007), et l'analyse discriminante linéaire (FAOUZI et COLLIOT 2023). Des techniques telles que la machine à vecteurs de support (VAPNIK 1963) et les modèles basés sur des arbres enrichissent également cette catégorie. D'autre part, l'apprentissage non supervisé offre des approches de regroupement comme l'algorithme des k-moyennes (FAOUZI et COLLIOT 2023) et les mixtures de gaussiennes (C. M. BISHOP 2007; FAOUZI et COLLIOT 2023), ainsi que des méthodes de réduction de dimensionnalité, telles que l'analyse en composantes principales (C. M. BISHOP 2007), l'analyse discriminante linéaire (FAOUZI et COLLIOT 2023), et l'analyse en composantes principales du noyau (FAOUZI et COLLIOT 2023).

Les approches non neuronales restent un outil fondamental de l'apprentissage automatique, offrant robustesse, efficacité et clarté là où les méthodes plus récentes pourraient ne pas être nécessaires ou appropriées (JAN et al. 2019). La décision d'utiliser des techniques non neuronales doit donc être guidée par la nature spécifique de la tâche, la quantité et la qualité des données disponibles, ainsi que par le besoin d'interprétabilité des résultats obtenus.

2.3.3 Les mélanges de lois de gaussiennes

Dans le champ de l'analyse de données, la mixture de gaussiennes est reconnue comme une méthode statistique puissante pour la modélisation des distributions complexes à travers la combinaison de plusieurs distributions normales. Ce modèle est essentiellement utilisé pour identifier des groupes inhérents au sein de données multidimensionnelles, permettant ainsi une meilleure compréhension et classification des observations. Dans cette sous-section, nous nous emploierons à rappeler les mixtures de gaussiennes, pour plus de détails, se référer à C. BISHOP (2006).

Dans le cas d'une variable aléatoire binaire z de K dimensions dans lequel l'élément particulier $z_k = 1$ et le reste est égal à 0, avec $z_k \in \{0, 1\}$ et $\sum_k z_k = 1$, on peut écrire la distribution marginale sur z en fonction du coefficient de mélange π_k :

$$p(z_k = 1) = \pi_k \quad (1)$$

avec $0 \leq \pi_k \leq 1$ et $\sum_{k=1}^K \pi_k = 1$.

Parce que z utilise 1-à-K représentation, nous pouvons écrire la distribution de la manière suivante :

$$p(z) = \prod_{k=1}^K \pi_k^{z_k} \quad (2)$$

De manière similaire, la distribution conditionnelle de x pour une valeur particulière de z est une gaussienne :

$$p(x|z_k = 1) = \mathcal{N}(x|\mu_k, \Sigma_k) \quad (3)$$

avec μ_k la moyenne et Σ_k la covariance. Elle peut aussi être écrite sous la forme :

$$p(x|z_k) = \prod_{k=1}^K \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k} \quad (4)$$

On définit la distribution d'un mélange gaussien comme la somme de la distribution jointe sur tous les états possibles de z :

$$p(x) = \sum_z p(z)p(x|z) = \sum_{k=1}^K \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \quad (5)$$

Une autre variable importante est la probabilité conditionnelle de z sachant x . Nous utilisons $\gamma(z_k)$ pour représenter $p(z_k = 1|x)$, dont la valeur peut être calculée en appliquant le théorème de Bayes :

$$\gamma(z_k) \equiv p(z_k = 1|x) = \frac{p(z_k = 1)p(x|z_k = 1)}{\sum_{j=1}^K p(z_j = 1)p(x|z_j = 1)} \quad (6)$$

$$= \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)} \quad (7)$$

La variable π_k peut être vue comme la probabilité a priori de $z_k = 1$, et la quantité $\gamma(z_k)$ comme la probabilité a posteriori une fois que la variable x a été observée. On appelle $\gamma(z_k)$ la responsabilité que le composant k prend pour expliquer l'observation x .

Supposons que nous ayons un jeu de données d'observations $\{x_1, \dots, x_N\}$. On peut représenter ces données comme une matrice \underline{X} de dimension $N \times D$, avec D la dimension des données d'observation x_n . Les variables latentes correspondantes sont représentées par une matrice \underline{Z} de dimension $N \times K$. En supposant que les points de données sont tirés indépendamment, on peut écrire la fonction logarithmique de vraisemblance de la manière suivante :

$$\ln p(\underline{X}|\pi, \mu, \Sigma) = \sum_{n=1}^N \ln \left[\sum_{k=1}^K \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right] \quad (8)$$

L'objectif est alors de trouver le paramètre θ qui maximise la vraisemblance, avec : $\theta = (\{\mu_j\}_{j=1}^K, \{\Sigma_j\}_{j=1}^K, \{\pi_j\}_{j=1}^K)$.

Afin de résoudre ce problème de maximisation, l'algorithme Expectation-Maximization (EM) (DEMPSTER, LAIRD et RUBIN 1977) est couramment utilisé. Après avoir initialisé les paramètres de chaque distribution, une boucle en deux étapes est répétée jusqu'à convergence, qui correspond à une stabilité des paramètres à travers les itérations successives :

- l'étape d'espérance, au cours de laquelle est calculée la probabilité que chaque échantillon $x(i)$ ait été généré par la distribution,
- l'étape de maximisation, où l'on met à jour la probabilité et les paramètres de chaque distribution.

Algorithmus 1 : Expectation-Maximization pour les Modèles de Mélange Gaussien

Entrée : Paramètres initiaux $\Theta = \{\mu_k, \Sigma_k, \pi_k\}$ pour $k = 1, \dots, K$
Sortie : Paramètres optimisés $\Theta = \{\mu_k, \Sigma_k, \pi_k\}$

```

1 while jusqu'à convergence do
  // E-step : Calculer la probabilité a posteriori  $\gamma_i(k)$  pour chaque  $i$ 
  // et  $k$ 
2  foreach  $i, k$  do
3     $\gamma_i(k) = \frac{\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(x_i | \mu_j, \Sigma_j)}$ 
  // M-step : Mettre à jour chaque paramètre
4  foreach  $k$  do
5     $\mu_k = \frac{\sum_{i=1}^n \gamma_i(k) x_i}{\sum_{i=1}^n \gamma_i(k)}$ 
6     $\Sigma_k = \frac{\sum_{i=1}^n \gamma_i(k) (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^n \gamma_i(k)}$ 
7     $\pi_k = \frac{\sum_{i=1}^n \gamma_i(k)}{n}$ 

```

Le modèle de mélange gaussien présente plusieurs avantages. Tout d'abord, l'utilisation des densités de distribution normale au lieu des distances euclidiennes pour la comparaison des données atténue le problème de dilatation dans les espaces de grande dimension. De plus, le modèle de mélange gaussien intègre des matrices de covariance, permettant la formation de clusters aux formes elliptiques.

Néanmoins, le modèle de mélange gaussien présente également plusieurs inconvénients. Le nombre de distributions k est un hyper-paramètre. Choisir une valeur très différente du nombre réel de clusters, ou modes de la distribution, peut entraîner la formation de clusters médiocres. Deuxièmement, la log-vraisemblance n'est pas une fonction concave. L'algorithme EM garantit la convergence, mais il peut converger vers un maximum local qui n'est pas un maximum global. Plusieurs stratégies sont souvent mises en œuvre pour résoudre ce problème, y compris une initialisation sophistiquée des centroïdes (ARTHUR et VASSILVITSKII 2007) et l'exécution de l'algorithme plusieurs fois tout en conservant la meilleure exécution (c'est-à-dire celle qui produit la log-vraisemblance la plus élevée).

2.3.4 Les approches par réseaux de neurones

Les réseaux de neurones en informatique sont des modèles inspirés par le fonctionnement du cerveau humain, utilisés pour reconnaître des schémas et résoudre divers problèmes complexes en utilisant un mode d'apprentissage principal, la descente de gradient (HECHT-NIELSEN 1989), dans des domaines scientifiques variés comme la vision par ordinateur, le traitement du langage naturel et la prédiction de séries temporelles et trouvent des applications dans la santé (SHAMSHIRBAND et al. 2021), la fabrication

intelligente (J. WANG et al. 2018), la cybersécurité (Priyanka DIXIT et SILAKARI 2021), la robotique (PIERSON et GASHLER 2017)...

Parmi les types les plus courants, on trouve les réseaux de neurones convolutifs (CNNs), qui excellent dans le traitement des images et sont fondamentalement détaillés dans l'article fondateur de LECUN et al. (1998), ils fonctionnent en appliquant des filtres convolutionnels successifs pour extraire des caractéristiques locales des données d'entrée, suivis de couches de regroupement et de couches entièrement connectées pour classer ou prédire les résultats basés sur ces caractéristiques extraites. On retrouve des architectures telles que Alexnet (KRIZHEVSKY, SUTSKEVER et HINTON 2012), VGG-16 (SIMONYAN et ZISSERMAN 2015) ou encore GoogleNet (SZEGEDY et al. 2015).

Les réseaux de neurones récurrents Recurrent Neural Network (RNN), sont des modèles connexionnistes qui capturent la dynamique des séquences grâce à des cycles dans le réseau de nœuds, ils sont particulièrement utiles pour les données séquentielles comme le texte ou le son. Les deux piliers de ce groupe de ce type de réseaux de neurones sont les Long Short-Term Memory (LSTM) (HOCHREITER et SCHMIDHUBER 1997) et les Bidirectional Recurrent Neural Network (BRNN) (SCHUSTER et PALIWAL 1997). En 2014, GRAVES, WAYNE et DANIHELKA (2014) améliorent les réseaux de neurones récurrents en leur ajoutant une mémoire externe adressable. Cette avancée permet aux RNN de mieux exécuter des tâches algorithmiques complexes comme le tri.

Les réseaux adverses génératifs (GANs), qui apprennent à générer de nouvelles données similaires aux données d'entraînement, ont été proposés par GOODFELLOW et al. (2020), révolutionnant la génération d'images et de vidéos. Le modèle de base du réseau adverse génératif est un jeu à deux réseaux et deux joueurs, avec un objectif à somme nulle basé sur la théorie des jeux. Le générateur, qui est entraîné sur un ensemble de données de vrais échantillons, tente de produire des échantillons convaincants capables de tromper le discriminateur, de sorte que ce dernier croit que les échantillons sont authentiques. Ce modèle est considéré comme un apprentissage semi-supervisé, et les poids sont ajustés par rétropropagation. Le jeu se termine lorsque le discriminateur ne peut distinguer les échantillons réels des échantillons générés qu'avec une précision de 50%, ce qui équivaut à faire un choix binaire ou un pile ou face.

L'arrivée des modèles de langage de grande taille Large Language Model (LLM) marque une évolution significative dans le domaine de l'intelligence artificielle. Ces modèles, tels que GPT (Generative Pre-trained Transformer) et BERT (Bidirectional Encoder Representations from Transformers) (DEVLIN et al. 2019), utilisent des réseaux de neurones profonds pour générer, comprendre et interpréter le langage humain de manière incroyablement nuancée. Le principe de ces modèles est fondé sur l'architecture des Transformers, qui permet une analyse efficace des séquences de données (comme les textes) grâce à des mécanismes d'attention (VASWANI et al. 2017) qui pondèrent l'importance relative des différents mots dans une phrase. Par exemple, GPT-3, avec ses 175 milliards de paramètres, montre comment l'augmentation de la taille du modèle peut entraîner une amélioration significative de la capacité à générer des textes cohérents et contextuellement appropriés (BROWN et al. 2020). Ces modèles de langage de grande taille ont transformé la compréhension du langage par les machines et ouvert de nouvelles perspectives pour des applications variées, telles que la traduction automatique, la création de contenu et l'assistance conversationnelle. Intégrés à des robots sociaux, ces agents conversationnels

peuvent enrichir les interactions humaines et faciliter la gestion de multiples situations, favorisant ainsi une intégration plus naturelle de la technologie dans notre quotidien.

2.4 LES SIGNAUX DE L'ENVIRONNEMENT

2.4.1 Reconnaissance automatique de la parole

La parole est le mode de communication le plus courant et naturel utilisé entre les humains. Il est raisonnable de supposer que les individus se sentent plus à l'aise en utilisant la parole comme moyen d'interaction avec diverses machines, plutôt que d'autres modes de communication tels que les claviers ou encore une tablette tactile. Un système de reconnaissance automatique de la parole permet de transformer le signal audio perçu en sa signification textuelle de la parole. Les premiers modèles de reconnaissance automatique de la parole se basent sur quatre étapes, l'extraction des composantes acoustiques, la modélisation acoustique, la modélisation du langage et la recherche basée sur des modèles de décision Bayésien (BAYES et PRICE 1763). L'essor de l'apprentissage profond tend progressivement à remplacer les méthodes statistiques telles que les mixtures de gaussiennes (BOURLARD et MORGAN 1993) ou les analyses discriminantes non linéaires (FONTAINE, RIS et BOITE 1997) par des modèles plus performants. Ainsi, des modèles classiques de reconnaissance automatique de la parole (ASR) utilisant des réseaux de neurones, comme les réseaux de neurones récurrents (GRAVES, FERNÁNDEZ et al. 2006), ont fait leur apparition, leur architecture séquentielle étant particulièrement adaptée à ce problème. C'est également dans ce domaine que les premiers modèles basés sur l'attention ont été testés (CHOROWSKI et al. 2015). Cependant, l'approche séquentielle n'est pas la seule utilisée en reconnaissance automatique de la parole, puisque des modèles de bout en bout ont aussi été adoptés (LIANG et al. 2006 ; GRAVES 2012 ; GRAVES et JAITLY 2014).

Les techniques plus récentes d'apprentissage automatique se sont révélées particulièrement efficaces. Par exemple, KRIMAN et al. (2019) présente QuartzNet, un modèle acoustique neuronal pour la reconnaissance automatique de la parole. Ce modèle utilise des convolutions séparables 1D pour atteindre une précision proche de l'état de l'art tout en réduisant le nombre de paramètres. Inspiré par l'architecture Jasper (J. LI et al. 2019), QuartzNet diminue la taille et les besoins en calcul du modèle, facilitant ainsi un entraînement et un déploiement rapides. De son côté, SAINATH et al. (2019) propose un modèle de reconnaissance automatique de la parole (ASR) en deux passes, combinant un transducteur RNN-T pour la première passe et un modèle "Listen", "Attend" and "Spell" (LAS) pour la seconde. Cette architecture permet de réduire le taux d'erreur de mots WER de 17 % à 22 % par rapport au modèle RNN-T seul, tout en maintenant une latence additionnelle faible, inférieure à 200 ms.

Cependant, l'arrivée des LLMs a considérablement changé le domaine de l'ASR, améliorant à la fois la précision et l'adaptabilité. Un exemple notable est Whisper d'OpenAI, un système ASR polyvalent entraîné sur un ensemble de données diversifiées multilingues et multitâches supervisées, permettant une performance robuste à travers diverses tâches de reconnaissance vocale (RADFORD et al. 2022).

La tâche de reconnaissance de la parole constitue l'un des modules indispensables à intégrer au projet μ DialBot, permettant à l'agent de comprendre les propos de ses interlocuteurs et d'y répondre de manière appropriée.

2.4.2 *Détection et suivi de personnes par vision*

La détection et le suivi humains sont de plus en plus sollicités dans diverses applications, allant des interfaces homme-machine à la robotique, l'analyse de foules, la surveillance, l'automatisation et les applications médicales. Plus spécifiquement en robotique, pour mettre en œuvre un système d'interaction humain-robot réussi, un robot doit posséder des capacités avancées pour réaliser des tâches complexes. Cela inclut la reconnaissance et le suivi précis des mouvements et des comportements humains, afin de permettre une interaction plus naturelle et efficace entre l'homme et la machine. La tâche de détection de personnes, aussi appelée détection d'objet, consiste à trouver toutes les instances des classes prédéfinies et de les localiser grossièrement dans l'image à l'aide de boîtes alignées sur les axes.

Les travaux pionniers dans la détection d'objets comprennent le détecteur Viola-Jones (VIOLA et JONES 2001), qui combine des caractéristiques telles que les fonctionnalités de type Haar, l'image intégrale, Adaboost et le classificateur en cascade. Le descripteur de caractéristiques Histogram of Oriented Gradients (HOG) (DALAL et TRIGGS 2005) a été proposé par Dalal et Triggs, améliorant la détection d'objets, notamment les piétons, en utilisant les gradients et leur orientation. Plus tard, le modèle de parties déformables (FELZENSZWALB, MCALLESTER et RAMANAN 2008) propose une utilisation des parties individuelles des objets pour une détection plus précise, surpassant HOG en précision.

L'arrivée du deep learning va marquer l'arrivée de nouveaux types de modèles. Un réseau comportant un module séparé pour générer des propositions de régions est appelé détecteur à deux étapes (R. GIRSHICK et al. 2014; HE, X. ZHANG et al. 2014; ROSS GIRSHICK 2015; T.-Y. LIN, DOLLÁR et al. 2017; DAI et al. 2016; HE, GKIOXARI et al. 2018). Ces modèles cherchent à identifier un nombre arbitraire de propositions d'objets dans une image lors de la première étape, puis les classifient et les localisent dans la seconde. Du fait de ces deux étapes distinctes, ces systèmes sont généralement plus lents pour générer des propositions, ont une architecture complexe et manquent de contexte global. En revanche, les détecteurs à une étape (REDMON et al. 2016; W. LIU et al. 2016; T.-Y. LIN, GOYAL et al. 2018; ZHOU, D. WANG et KRÄHENBÜHL 2019) classifient et localisent les objets sémantiques en une seule fois en utilisant un échantillonnage dense. Ils utilisent des boîtes ou des points clés prédéfinis de différentes échelles et proportions pour localiser les objets, offrant ainsi une performance en temps réel supérieure et une conception plus simple par rapport aux détecteurs à deux étapes.

2.4.3 *Diarisation du locuteur*

L'association du signal vocal à sa source d'émission est une tâche cruciale pour réaliser une analyse précise d'une scène comportant différents intervenants. Communément appelée diarisation des locuteurs, elle consiste à attribuer des segments audio à des classes

correspondant aux identités des locuteurs. Cette tâche répond à la question « qui a parlé et quand ? » (TRANter et D.A. REYNOLDS 2006).

Les premiers travaux sur la diarisation des locuteurs se sont concentrés sur la modalité audio (HERSHEY et MOVELLAN 1999; D.A. REYNOLDS et TORRES-CARRASQUILLO 2005). Aujourd'hui, les systèmes typiques de diarisation des locuteurs audio utilisant l'audio comme entrée se composent de trois composants : (1) la segmentation de la parole EL-KHOURY, SENAC et PINQUIER 2009, où l'entrée audio est décomposée en courts segments, chaque segment étant censé ne contenir qu'un seul locuteur, et le bruit est filtré. Ce composant peut être considéré comme un module de détection d'activité vocale. (2) L'extraction de l'incorporation audio des sections segmentées à travers diverses méthodes, parmi lesquelles les plus notables sont les MFCCs (Mel-frequency Cepstral Coefficients, KENNY, Douglas REYNOLDS et CASTALDO 2010), les facteurs de locuteurs (CASTALDO et al. 2008) et les i-vectors (SHUM et al. 2013). Et (3) un module de regroupement, où les incorporations audio extraites sont classées en locuteurs. Pour cette tâche, le nombre de locuteurs est déterminé.

Il est également possible de pré-traiter en amont en utilisant des techniques d'amélioration de la parole et de débruitage, menant à une amélioration significative grâce à l'apprentissage profond (GAO et al. 2018). La représentation des locuteurs a été grandement améliorée avec l'arrivée des réseaux neuronaux et d'autres nouvelles méthodes telles que le d-vector (Q. WANG et al. 2017) et le x-vector (VILLALBA et al. 2019).

Une alternative intéressante consiste en la fusion des données audio et visuelles. La combinaison de ces deux modalités fournit des informations complémentaires et, par conséquent, les approches audiovisuelles de la diarisation des locuteurs sont susceptibles d'être plus robustes que les approches uniquement audio. Elle peut être associée à un suivi du visage ou de la bouche grâce à une coïncidence spatiale sur le plan de l'image (GEBRU, Silèye BA et al. 2015; GEBRU, Sileye BA et al. 2018), sur un plan au sol (Y. WANG et CAVALLARO 2016) ou en 3D (XIONG et al. 2021).

Les méthodes de suivi des personnes en 3D nécessitent des réseaux de caméras et de microphones répartis dans l'espace, ce qui n'est pas praticable dans le cas de scénarios complexes (réels). D'un autre côté, les méthodes reposant sur l'image plane ou au sol peuvent manquer d'informations et souffrir davantage d'occlusions, mais elles offrent l'avantage d'une collecte de données plus aisée et peuvent être utilisées dans une gamme plus large de scénarios.

En plus de la position de la bouche, plusieurs méthodes sont basées sur la synergie entre les énoncés et les mouvements des lèvres à travers différentes approches telles que l'information mutuelle (HERSHEY et MOVELLAN 1999) et l'apprentissage profond (CHUNG, B.-J. LEE et I. HAN 2019).

Plus tard dans le manuscrit, nous proposons une méthode modélisant cette fusion à travers la coïncidence spatiale de la localisation des sources visuelles et sonores SSL et en combinant ce modèle de concordance avec une formulation bayésienne dynamique qui suit l'identité du locuteur actif. Le SSL offre plusieurs avantages dans les conversations multi-utilisateurs, tels que la capacité de gérer les segments de parole se chevauchant, éliminant le besoin d'un module d'activation vocale. La méthode proposée peut être appliquée dans diverses conditions acoustiques en tirant parti des informations spatiales du SSL et de la localisation du visage.

2.4.4 *Les signaux sociaux*

Le cœur de la connexion humaine ne se trouve pas toujours dans nos paroles, mais plutôt dans les nuances discrètes et non dites qui façonnent nos échanges. Bien que cet aspect essentiel de la communication soit souvent ignoré, il est crucial pour bâtir des relations, partager des émotions et créer une dynamique sociale.

Ces comportements non verbaux sont intrinsèquement ambigus et ne sont pas toujours liés à des significations spécifiques, car leur interprétation peut varier en fonction du contexte, de la culture et de facteurs non liés (KNAPP et HALL 1972). Par exemple, les postures et les distances physiques peuvent refléter des attitudes sociales ou simplement le confort et les contraintes physiques (SCHEFLEN 1964).

Cette communication se fait à travers une série complexe de gestes et expressions non verbales qui, en toute discrétion, dirigent nos interactions. Elle va au-delà des mots et des différences culturelles, nous permettant de comprendre les intentions et les émotions cachées derrière ce que nous disons. Ces gestes et expressions sont ce que nous appelons les signaux sociaux. Le traitement des signaux sociaux consiste à créer des modèles, analyser et synthétiser les comportements et signaux émis par les individus lors d'interactions, que ce soit entre humains ou entre humains et machines, selon VINCIARELLI, SALAMIN et M. PANTIC (2009). Dans leur étude, VINCIARELLI, Maja PANTIC et al. (2008) élaborent une taxonomie des indices comportementaux non verbaux qui sont liés à des signaux sociaux importants. Ces indices comportementaux sont regroupés en cinq catégories distinctes et représentées dans la table 2.

Cette classification détaillée aide à mieux comprendre comment les comportements non verbaux peuvent être interprétés et utilisés pour analyser les interactions sociales dans divers contextes.

L'utilisation des signaux sociaux joue un rôle important dans les interactions entre humains et robots, car ces signaux induisent des comportements distincts dans les interactions humain-robot par rapport aux interactions entre humains. KALASHNIKOVA et al. (2023) analysent la fréquence fondamentale moyenne (F_0) et la variance de F_0 dans la parole dirigée vers un humain, un robot ou une enceinte, chez des participants masculins et féminins. Ils mettent en évidence que le degré d'humanité de l'interlocuteur, allant du plus humain (la personne) au moins humain (l'enceinte domestique), influence l'intonation de la voix. SKANTZE, HJALMARSSON et OERTEL (2014) ont démontré que l'incertitude des participants lors d'une tâche de création de carte entre un humain et un robot se manifestait par des variations prosodiques et des choix lexicaux spécifiques. De plus, HOUGH et SCHLANGEN (2017) ont proposé un modèle de "grounding" pour l'interaction humain-robot, où le robot exprimait son incertitude par des mouvements plus lents, perçus par les utilisateurs comme des signaux d'incertitude. Par ailleurs, MEHLMANN et al. (2014) ont montré que les auditeurs utilisent le regard pour indiquer leur compréhension de l'objet auquel l'orateur fait référence, en lien avec le processus physiologique de localisation de cet objet. On peut également noter la présence des émotions (S. LI et DENG 2022; CLADIÈRE et al. 2023), qui apportent une dimension sociale significative lors d'une interaction, par exemple.

En termes d'engagement, le regard mutuel peut être utilisé comme un signe indiquant qu'un utilisateur souhaite interagir avec le système. KUNO et al. (2007) ont découvert que

Indices sociaux	Exemples de signaux sociaux						
	émotion	personnalité	statut	domination	persuasion	régulation	rapport
Apparence physique							
taille		✓	✓				
attractivité	✓	✓	✓				
forme du corps		✓	✓				
Gestes et posture							
gestes de la main				✓	✓		✓
posture		✓	✓	✓	✓		✓
marche			✓	✓	✓		
Comportement du visage et des yeux							
expressions faciales	✓	✓					✓
comportement du regard		✓	✓	✓		✓	✓
focus de l'attention	✓		✓	✓	✓		✓
Comportement vocal							
prosodie	✓	✓		✓	✓		✓
tour de parole			✓		✓	✓	
vocalisations	✓		✓	✓	✓		
silence	✓		✓			✓	
Espace et Environnement							
distance	✓	✓	✓	✓	✓		✓
disposition des sièges		✓		✓			

TABLE 2 – Tableau des indices comportementaux associés à certains des comportements sociaux les plus importants VINCIARELLI, Maja PANTIC et al. 2008.

le regard mutuel et les hochements de tête concomitants étaient des indicateurs importants de l'engagement du public lors de la présentation d'un robot dans un musée. De plus, NAKANO et ISHII (2010) ont présenté un modèle de regard comme signe d'engagement mutuel. Un agent utilisant un modèle de détection du regard plus sophistiqué a été jugé préférable par les participants aux tests.

Ainsi la communication humaine va bien au-delà des mots, reposant fortement sur des signaux non verbaux pour transmettre des émotions et des intentions. Ces indices comportementaux, bien que souvent subtils et dépendants du contexte culturel et situationnel, sont essentiels pour comprendre et analyser les interactions sociales. Les recherches dans le domaine des signaux sociaux, notamment en interaction humain-robot, montrent comment ces signaux peuvent être interprétés et utilisés pour améliorer la compréhens-

sion et l'engagement dans diverses situations. Ainsi l'utilisation de ces signaux dans le cadre du projet μ DialBot doivent permettre d'enrichir les relations et interactions avec des agents robotiques.

2.5 LA PROACTIVITÉ

2.5.1 *Dans la psychologie organisationnelle*

Dans le cadre du comportement humain et de la psychologie organisationnelle, d'où la proactivité tire la majorité de ses définitions, la proactivité est définie comme un comportement anticipatif, axé sur le changement et auto-initié, particulièrement manifeste dans les contextes professionnels BATEMAN et CRANT (1993) ou encore comme une action anticipative et autonome conçue pour influencer les individus et/ou leur environnement GRANT et ASHFORD (2008).

Ce comportement est typique des individus qui non seulement prennent l'initiative, mais anticipent également les défis potentiels, prenant des mesures préventives pour les atténuer ou pour capitaliser sur les opportunités émergentes. La proactivité va au-delà de simples réponses réactives aux stimuli environnementaux ; elle incarne une stratégie prospective marquée par un engagement actif et une planification méticuleuse. Ce paradigme a été rigoureusement examiné dans le domaine du comportement organisationnel, révélant son association avec de nombreux résultats positifs, dont une performance professionnelle améliorée, une satisfaction professionnelle accrue et un bien-être global (PARKER, BINDL et STRAUSS 2010). Les sciences du comportement ont postulé cette capacité comme une marque évolutive exclusive à Homo sapiens, facilitant notre capacité à s'engager à la fois dans des comportements collaboratifs et proactifs (TOMASELLO et al. 2005).

2.5.2 *Dans la HRI*

La littérature sur la proactivité dans la HRI reprend les concepts posés par la psychologie organisationnelle du comportement proactif. Selon RASHED et al. (2018), une approche HRI proactive pour les robots sociaux est caractérisée par une action anticipatoire initiée par le robot pour avoir un impact sur lui-même ou sur les autres. BARAGLIA et al. (2017) définit le terme d'aide proactive initiée par un robot comme une aide du robot dès qu'elle est possible, contrairement à l'aide réactive initiée par le robot dans laquelle l'assistance est enclenchée lorsque le robot détecte que l'utilisateur a besoin d'aide. Cette définition peut manquer de clarté car dès lors qu'un robot social détecte une personne enrhumée et lui apporte un mouchoir, le comportement peut être vu comme proactif et l'on a bien une aide enclenchée lorsque le robot détecte que l'utilisateur a besoin d'aide. Si l'aide demandée ne provient pas explicitement de l'utilisateur, on aura souvent plus souvent affaire à de la réactivité. La frontière étymologique entre le caractère proactif et réactif de certaines situations est fine. SIRITHUNGE, JAYASEKARA et D. PATHIRANA (2019) complète cette définition en qualifiant de robot proactif une entité qui discerne les exigences d'une situation donnée et agit rapidement sans directives extérieures. C.

et CHODOROWSKI (2019) proposera une définition simplifiée basée sur cette absence de demande : agir sans sollicitation.

Pour les besoins de cette recherche, nous nous alignons sur la définition de la proactivité la plus complète centrée sur l'homme : *la capacité d'initier de manière autonome des actions anticipatoires, fondées sur le raisonnement, visant à influencer les individus et/ou leur environnement*, telle qu'elle est décrite par GROSINGER (2022). Jasmin Grosinger poussera la définition en classant la proactivité selon trois catégories :

- Le premier type est la *proactivité pour soutenir l'intention humaine*. Ce type de proactivité consiste à aider l'humain à atteindre son intention ou objectif par des actions anticipatrices et auto-initiées. Par exemple, cela inclut la reconnaissance de l'intention humaine et l'assistance proactive pour réaliser cette intention, comme un robot domestique qui prédit les actions suivantes de l'humain pour les aider dans une tâche.
- Le deuxième type est la *proactivité avec un objectif donné*. Ce type de proactivité se manifeste lorsqu'un objectif explicite est donné à l'agent (par l'humain ou un déclencheur externe). L'agent réalise des actions proactives pour atteindre cet objectif spécifique. Par exemple, un système robotique avec une couche éthique peut simuler différentes alternatives de comportement pour choisir et exécuter l'action la plus éthique en fonction des objectifs externes fournis. Il est le type de proactivité qui correspond le mieux à la proactivité utilisée dans cette thèse, l'objectif du robot et d'interagir proactivement avec les patients dans une salle.
- Le troisième type est la *proactivité à partir de principes fondamentaux*. Ce type de proactivité est basé sur le raisonnement à partir de principes fondamentaux. Il vise à comprendre les facteurs et les capacités cognitives nécessaires pour générer un comportement proactif, ainsi que leurs interactions. Un exemple de cette approche est l'utilisation de modèles probabilistes pour informer les décisions d'action d'un robot social adaptatif, en apprenant sur le terrain l'impact des actions du robot sur l'utilisateur.

2.5.3 Modélisation HRI de la pro-activité

2.5.3.1 Approches générales de modèles proactifs

La pro-activité dans notre société peut s'exprimer sous de nombreuses formes. La reproduction de ce mécanisme comportemental dans tout ce qu'il englobe en HRI est une tâche complexe, malgré tout, il existe des approches générales de modélisation de la proactivité. GROSINGER, PECORA et SAFFIOTTI (2016), GROSINGER, PECORA et SAFFIOTTI (2017) et GROSINGER, PECORA et SAFFIOTTI (2018) modélisent la proactivité comme une question de maintien de l'équilibre, c'est-à-dire la recherche d'opportunités d'action permettant de maintenir le système dans des états souhaitables tout en évitant les états indésirables. La désirabilité des états est définie en tant que préférences graduées, en employant des mécanismes issus du domaine de la logique floue. Le modèle de formalisation croyance-désir-intention (McCANN et BRATMAN 1991) est également une approche assez répandue dans la modélisation de la proactivité (YORKE-SMITH et al. 2012; BREMNER et al. 2019a; C. et CHODOROWSKI 2019). L'approche Belief Desire Intention (BDI)

est un modèle de programmation des agents qui s’inspire de la théorie philosophique de l’action humaine. Ce modèle a été conçu pour simuler le raisonnement pratique des agents intelligents et leur prise de décision basée sur des modèles cognitifs humains. Les croyances représentent l’information que l’agent possède sur le monde, y compris sur lui-même et sur les autres agents, les désirs sont les objectifs ou les états que l’agent souhaite atteindre et les intentions sont des plans d’action que l’agent s’engage à exécuter pour réaliser ses désirs.

2.5.3.2 *Engagement proactif*

Au cours des dernières années, la recherche en interaction humain-robot a considérablement progressé, notamment dans la conception et l’évaluation de stratégies d’interactions proactives visant à améliorer l’efficacité et le naturel de ces interactions. Par exemple, KOIDE et al. (2004) ont développé une méthode de guidage proactif basée sur des règles, exploitant des capteurs ubiquitaires dans une salle d’exposition. D’autres modèles d’approche proactive ont été introduits pour divers contextes comme l’engagement de personnes en mouvement (SATAKE et al. 2009) ou encore de visiteurs de musées (YOUSUF et al. 2013).

KANDA, GLAS et al. (2010) présentent un modèle proactif pour les robots sociaux opérant dans des espaces publics tels que les centres commerciaux, en utilisant des techniques d’abstraction pour analyser les trajectoires. Il définit les comportements locaux tels que des comportements basiques et primaires, comme la marche ou la course. Les comportements globaux représentent des comportements plus complexes et avancés, tels que les trajectoires générales composées de plusieurs comportements locaux en séquence. Des expériences sur le terrain valident que ce cadre permet au robot d’approcher proactivement et de servir efficacement les clients potentiels en prédisant leurs emplacements et comportements probables quelques secondes à l’avance.

SHIOMI, KANDA et al. (2010) ont étendu cette réflexion en intégrant un comportement proactif dans les tâches de guidage touristique effectuées par des robots sociaux, avec un intérêt particulier pour encourager la participation des personnes qui n’interagissent pas directement avec le robot par le biais de mouvements avant et arrière. Leurs résultats expérimentaux confirment que cette approche encourage effectivement une plus grande écoute de la part des personnes présentes. KATO, KANDA et ISHIGURO (2015) ont exploré comment les robots peuvent approcher les humains de manière à la fois polie et socialement appropriée, ajoutant une autre dimension à la recherche sur les interactions proactives.

A partir d’un cadre proactif de robot de service, PENG et al. (2019) généralise son processus d’aide à la décision en 3 différents types de proactivité avec pour chacune quatre étapes, l’initiation, élicitation des préférences et recommandation, justification, et recherche du ressenti utilisateur. L’initiation de l’interaction se fait lorsque le robot détecte un besoin d’assistance dans le cas d’une proactivité importante et modérée. Le niveau de faible proactivité correspond à une requête utilisateur, mais dans ce cas, on parlera plutôt de comportement réactif. Néanmoins, ces niveaux de proactivité sont très dépendants du contexte.

Les travaux cités mettent en lumière l'importance des stratégies d'interaction proactive en HRI. Les avancées dans ce domaine, telles que le guidage proactif basé sur des règles, l'engagement de personnes en mouvement et la facilitation des interactions dans des lieux publics, témoignent de la diversité des approches envisageables. La recherche a également souligné l'importance de l'adaptabilité des robots en fonction des contextes, en intégrant des niveaux de proactivité variables et en tenant compte de l'aspect social et éthique des interactions. Ces développements, qui s'articulent autour de l'anticipation des besoins et comportements humains, ouvrent la voie à des interactions plus naturelles, efficaces et personnalisées entre les humains et les robots, marquant un pas significatif vers une intégration des robots dans la société.

2.5.3.3 *Analyse de la possibilité d'engagement proactif*

Dans le contexte de l'initiation proactive d'interactions, le système neurocognitif humain effectue une évaluation rapide de la probabilité de réussir une interaction avec un interlocuteur potentiel. Cette évaluation prend en compte divers facteurs, tels que le degré de familiarité de l'individu et son niveau actuel d'engagement dans d'autres activités ou interactions. Ce processus d'évaluation de la disponibilité interactionnelle permet de prendre une décision informée quant à l'opportunité d'initier une interaction ou de différer jusqu'à un moment plus propice. Il est à noter que le comportement non verbal constitue une part substantielle de cette évaluation, représentant approximativement 60 à 65% de la communication interpersonnelle totale, comme l'indique (BURGOON (1994)).

SIRITHUNGE, JAYASEKARA et D. PATHIRANA (2019) ont formalisé ce concept en tant que "mesure d'aptitude à l'interaction", définie comme le degré de prédisposition d'un individu à s'engager dans une interaction, mesuré à partir des signaux comportementaux observables émis par l'individu. Par ailleurs, Das et al. (2015) ont introduit le concept de "niveau d'attention visuelle" comme un paramètre quantifiable pour aider les agents autonomes à décider du moment optimal pour initier une interaction. Ce paramètre est modélisé en utilisant des schémas de regard et des indices contextuels. En complément, WEBB, GIULIANI et LEMAIGNAN (2022) ont développé une métrique d'engagement social visuel, validée à travers des simulations d'interactions sociales de groupe. Cette métrique est dérivée de deux signaux sociaux clés : la proxémique, qui concerne le positionnement spatial et sa gestion dans un contexte social, et le regard mutuel. Ces recherches mettent en avant le regard comme signal social dans les dynamiques sous-jacentes à l'interaction humain-robot.

2.6 PROJET μ DIALBOT

2.6.1 *Présentation*

Le projet Agence Nationale de Recherche (ANR) μ DialBot est un projet de l'Agence Nationale de Recherche française débuté en 2020 sur quatre années et rassemble 5 partenaires :

- Laboratoire Informatique d'Avignon, Avignon
- Lusage Living lab de l'hôpital Broca, Paris

- Le laboratoire Hubert Curien, Saint Etienne
- Institut national de recherche en sciences et technologies du numérique, Equipe RobotLearn, Grenoble
- ERM Automatismes, Carpentras

Les buts fondamentaux de ce projet résident dans la progression des méthodes statistiques et des algorithmes d'apprentissage profond, axés sur la fusion de données, l'intégration sensori-motrice, l'interprétation du comportement humain et la communication multimodale. L'objectif principal est d'implémenter ces composants dans une architecture robotique conçue pour fournir une assistance proactive aux individus âgés.

Le système doit faire l'objet d'une validation empirique dans les zones d'attente animées de l'hôpital Broca, regorgeant de patients, de leurs accompagnateurs et du personnel médical. Le rôle principal du robot est d'initier des conversations pour aider les patients et soutenir les soignants. Le robot doit agir de manière appropriée et rapide pour sélectionner et approcher un groupe ou une personne, initiant un dialogue. Le projet met l'accent sur les aspects éthiques, de sécurité et de confidentialité de la gestion des données. Le consortium s'engage à respecter scrupuleusement les réglementations en vigueur concernant le stockage, l'accès, l'utilisation et les protocoles de diffusion des données. La plateforme robotique prévue pour ce projet était à l'origine le robot ARI de Pal Robotics, mais du fait de retard de livraison le projet s'est finalement effectué sur la plateforme Pepper de United Robotics Group.

2.6.2 Organisation

Le projet μ DialBot est organisé en 4 lots de travail (figure 8) répartis sur les 4 années du projet. Les différents lots de travail peuvent être résumés de la manière suivante :

- **WP1** : développer des méthodes et des algorithmes pour extraire des indices de Human Behavior Understanding (HBU) à partir de données audio et visuelles. Concevoir des méthodes d'extraction améliorées qui sont robustes contre diverses perturbations dans un contexte réel et qui peuvent fournir une estimation quantitative de la fiabilité des indices extraits. Développer des algorithmes pour la reconnaissance à longue distance des activités individuelles et de groupe ainsi que pour l'estimation à courte distance des expressions faciales individuelles. Développer des méthodes et des algorithmes pour l'apprentissage en ligne des comportements des robots, conduisant à une extraction robuste des indices de HBU à courte distance nécessaires pour la communication en face-à-face (WP2).
- **WP2** : développer la capacité d'interaction vocale naturelle du robot. En plus de cela, un processus de prise de décision multimodal est développé permettant de combiner toutes les observations d'événements verbaux et non-verbaux avec des caractéristiques contextuelles pour définir une stratégie comportementale globale. La stratégie sera principalement apprise in situ, au moyen d'une procédure d'entraînement en ligne, et conditionnera les schémas d'engagement du robot dans ses différentes situations d'utilisation (multiparty, face-à-face).
- **WP3** : définir le protocole expérimental, spécifier les protocoles expérimentaux en laboratoire et en conditions réelles, et mener des expériences de type "Wizard of Oz". Intégrer progressivement les modules logiciels des WP1 et WP2 et mener des

expériences d'interaction dans les laboratoires. Évaluer l'architecture globale dans un contexte clinique avec de petits groupes de patients avant les expériences finales dans la salle d'attente.

- **WP4** : diffuser les résultats scientifiques du projet par le biais d'articles de journaux, de publications de conférences et de tutoriels ; présenter les principes, méthodes et résultats du projet lors d'ateliers spécialisés ; aborder la gestion scientifique, administrative et financière.

Cette thèse était centrale pour les objectifs et les problématiques portés par le projet μ DialBot dans son ensemble, et très intégrée dans son architecture complète. Les travaux de la thèse font partie intégrante du WP2 avec la mise en place de l'architecture HRI utilisée pour le projet, mais ils sont également fortement impliqués dans le WP1 avec le développement des modules ROS nécessaires à l'extraction des perceptions, en particulier la diarisation des locuteurs. Enfin, ils participent aussi au WP3 à travers l'intégration des modules logiciels WP1 et WP2, ainsi que dans la mise en place des protocoles expérimentaux et la conduite des expériences.

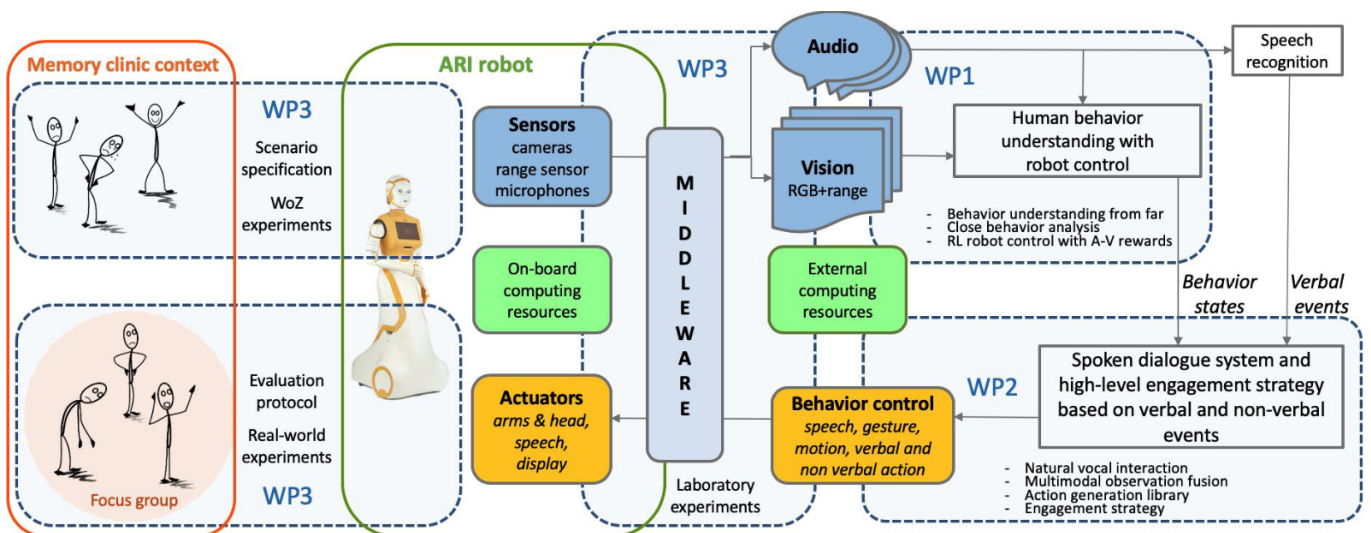


FIGURE 8 – Diagramme de l'organisation des lots de travail chargés de la conception du système de HRI dans projet μ DialBot.

Le WP2 est une composante du projet visant à développer un système de dialogue vocal et une stratégie d'engagement de haut niveau pour un robot, en se basant sur des événements verbaux et non verbaux. L'objectif est de créer une stratégie comportementale optimale grâce à un processus décisionnel multimodal qui combine l'estimation d'états verbaux et non verbaux ainsi que des caractéristiques contextuelles. Cette stratégie sera apprise sur place grâce à une procédure d'entraînement en ligne par renforcement profond.

Le système de dialogue sera construit sur les dernières avancées dans le domaine pour améliorer la naturalité, en utilisant des modèles innovants et un nouveau processus de développement. Il intégrera également un apprentissage en ligne et une adaptation à long terme pour gérer des utilisateurs non-experts. Une fusion de modalités sera établie pour connecter le système de dialogue avec des caractéristiques comportementales de

l'utilisateur, en utilisant un processus décisionnel multimodal. Ce processus devra choisir les meilleures actions multimodales, impliquant des réponses verbales et non verbales, et définira une stratégie d'engagement de haut niveau basée sur un planificateur de tâches hiérarchiques.

Pour optimiser cette stratégie d'engagement, il sera nécessaire de définir une fonction de récompense dans un cadre d'entraînement par renforcement. Différentes méthodes dans le domaine du renforcement seront explorées pour proposer des solutions efficaces. Une bibliothèque de haut niveau sera conçue pour interfacier les actions résultant de la stratégie d'engagement avec les actions réelles du robot. Des modules logiciels développés seront intégrés dans la plateforme ARI, avec un middleware léger compatible avec ROS, visant à développer des modules indépendants des détails techniques de la plateforme robotique et fonctionnels dans tout environnement informatique. Des expériences en laboratoire compléteront les étapes d'intégration logicielle.

2.6.3 *Impact et bénéfice du projet*

L'Agence Nationale de Recherche Française promeut le développement de systèmes robotiques autonomes et interactifs grâce à l'intelligence artificielle, μ DialBot s'alignant sur l'Axe 5.5 "Interaction, robotique" du Domaine 5 "Sciences du numérique". Ainsi, le projet vise à améliorer les robots humanoïdes avec des compétences d'interaction naturelle pour des dialogues situés en multi-parties, contribuant à l'autonomie conversationnelle et à la collaboration humain-robot dans les environnements de soins de santé. Le projet se concentre sur des résultats éthiques et mène des recherches interdisciplinaires, en particulier dans l'interaction humain-robot, en développant des modèles basés sur les données pour l'apprentissage à long terme et l'adaptation en fonction des personnalités des utilisateurs.

La technologie centrale du projet vise à augmenter l'acceptabilité des robots sociaux au-delà des soins de santé, en s'étendant à des secteurs tels que l'accueil dans les musées, la vente au détail, le divertissement et l'éducation. L'impact économique prévu est significatif, bénéficiant au marché de la technologie d'assistance et à l'industrie française. Les droits de propriété intellectuelle seront abordés, et le projet publiera des logiciels et des données en tant que ressources open-source pour bénéficier à la communauté de recherche.

Dans le domaine des soins de santé, μ DialBot améliorera la connaissance sur l'utilisation des robots sociaux d'assistance, en établissant des lignes directrices pour leur mise en œuvre et en améliorant la qualité des soins en automatisant des tâches et en réduisant la charge de travail des professionnels. Le projet explorera également d'autres scénarios d'application, tels que les écoles et les entreprises. Sur le plan académique, μ DialBot vise à accroître l'impact des publications dans des domaines tels que la vision par ordinateur, le traitement audio/parole et l'interaction homme-machine, et publiera sur des plateformes en libre accès.

Troisième partie

Contributions

CONTRIBUTIONS

3.1 INTRODUCTION

3.1.1 *Contexte*

L'objectif concret de cette thèse et du projet μ DialBot dans lequel elle s'inscrit est de concevoir un système multimodal proactif capable d'interagir vocalement avec les patients d'une salle d'attente d'un hôpital. Ce système doit être capable de percevoir la réalité à l'aide des capteurs de l'agent, d'interpréter la scène en fonction de ses perceptions et de prendre des décisions pour exécuter les actions adaptées. L'ensemble sera installé sur le robot humanoïde Pepper, dont les capacités de déplacement, de communication et la variété de capteurs font un choix idéal pour cette étude en interaction humain-robot. Une phase de tests sera ensuite menée pour valider le bon fonctionnement du système.

Dans ce chapitre, nous présenterons en détail le processus de conception d'un tel robot proactif multimodal. Cette conception repose sur l'utilisation de ROS, un outil essentiel pour gérer le flux constant des données entrantes (perceptions) et sortantes (actions) à travers les topics et pour structurer le dispositif en modules fonctionnels distincts. L'environnement fourni par ROS facilite non seulement la communication entre les différentes parties du système, mais il permet également une intégration flexible des capteurs et des actionneurs du robot. Grâce à son architecture modulaire, ROS permet une mise à jour et une maintenance plus aisées, tout en offrant une grande capacité d'extension pour de futures améliorations ou modifications du système, y compris son application sur un tout autre robot.

L'ensemble de la conception sera régi par une contrainte. Les modules développés doivent prendre en compte un ratio latence et mémoire vidéo afin que le système puisse être au maximum transportable sur un ordinateur portable disposant de 32 Go de RAM et doté d'une Quadro RTX 5000 mobile / Max-Q avec 16 Go de mémoire vidéo.

3.1.2 *Introduction à l'architecture FlowAct*

L'évolution des systèmes autonomes dans un contexte d'interaction humain-robot nécessite une synergie entre la perception continue de l'environnement et les actions potentielles pour naviguer ou interagir à l'intérieur de celui-ci. FlowAct est une architecture proactive d'interaction humain-robot multimodale, fonctionnant comme une boucle sans fin asynchrone de perceptions et d'actions et organisée par deux contrôleurs (Figure 9), le suivi de l'état de l'environnement et le planificateur d'actions. L'Environment State Tracker (EST) recueille et publie en permanence une représentation de l'environnement opérationnel, assurant ainsi un flux constant de données perceptives. Ce flux perceptif persistant est essentiel pour notre planificateur d'action avancé qui orchestre une collection de sous-systèmes d'action modulaires, tels que les modules de mouvement et

de parole, régissant leur déclenchement ou leur arrêt en fonction de l'évolution de la narration environnementale. L'EST utilise une fusion de diverses modalités sensorielles pour construire une représentation riche et en temps réel de l'environnement, qui est diffusée au planificateur d'action. Ce planificateur utilise un cadre décisionnel pour coordonner dynamiquement les modules d'action, ce qui leur permet de réagir de manière proactive et cohérente aux changements de l'environnement. Grâce à une série d'expériences réelles, nous démontrons l'efficacité de notre système dans le maintien d'une boucle perception-action continue, améliorant considérablement la réactivité et l'adaptabilité des agents autonomes. L'architecture modulaire des sous-systèmes d'action facilite l'extensibilité et l'adaptabilité à un large éventail de tâches et de scénarios. L'architecture FlowAct sera présentée en détail dans la section 3.4.

Plus précisément, dans ce chapitre, nous décrivons le flux continu de perceptions, en commençant par la collecte des *données brutes* directement à partir des capteurs de Pepper. Ces données sont ensuite analysées par des modules spécialisés pour être transformées en *perceptions raffinées*, qui serviront à construire une représentation détaillée de l'environnement. Nous détaillerons également la réalisation complète d'un module de *diarisation multimodal* basé sur l'image et le son, indispensable pour la compréhension de la scène. Ensuite, nous examinerons en détail l'*architecture du modèle proactif d'interaction humain-robot* que nous proposons, nommé FlowAct. Cela comprendra une description approfondie du module de création de l'environnement ainsi qu'une présentation du processus de prise de décision, orchestré par un module de planification d'action qui coordonne divers sous-modules d'action. Puis, dans un quatrième temps, nous décrivons les *expériences réalisées* en conditions naturelles pour valider l'ensemble de l'approche proposée.

Le choix de cette séquence de présentation, suivant l'ordre logique du traitement de l'information dans notre système, nous conduit à présenter de manière séquentielle dans les sections suivantes les aspects très techniques et pratiques du travail et les contributions scientifiques qui ont permis d'aboutir à notre objectif principal, le démonstrateur robotique interactif proactif.

3.2 LE FLOT DE PERCEPTION

3.2.1 Extraction des perceptions brutes

L'objectif de l'extraction des perceptions qui proviennent directement des capteurs du robot, consiste à récupérer ces données brutes, puis à les transmettre vers des topics ROS, sous la forme de messages bien formés et mis à la disposition des modules suivants du système. Nous l'avons vu dans la partie précédente, le robot multimodal Pepper est équipé de divers capteurs, notamment des caméras, des microphones, des sonars, des lasers et des gyroscopes. Il fonctionne sous NAOqi OS, un système d'exploitation basé sur Unix développé par Aldebaran. NAOqi permet de récupérer les données des capteurs en se connectant au robot par sockets réseaux. Il existe actuellement deux versions principales de NAOqi : la 2.5 et la 2.9. La version 2.5, la plus ancienne, permet de programmer Pepper en utilisant C++ et Python. En revanche, la version 2.9, plus récente, se base sur la bibliothèque Java. La programmation de modules de recherche étant plus simple en

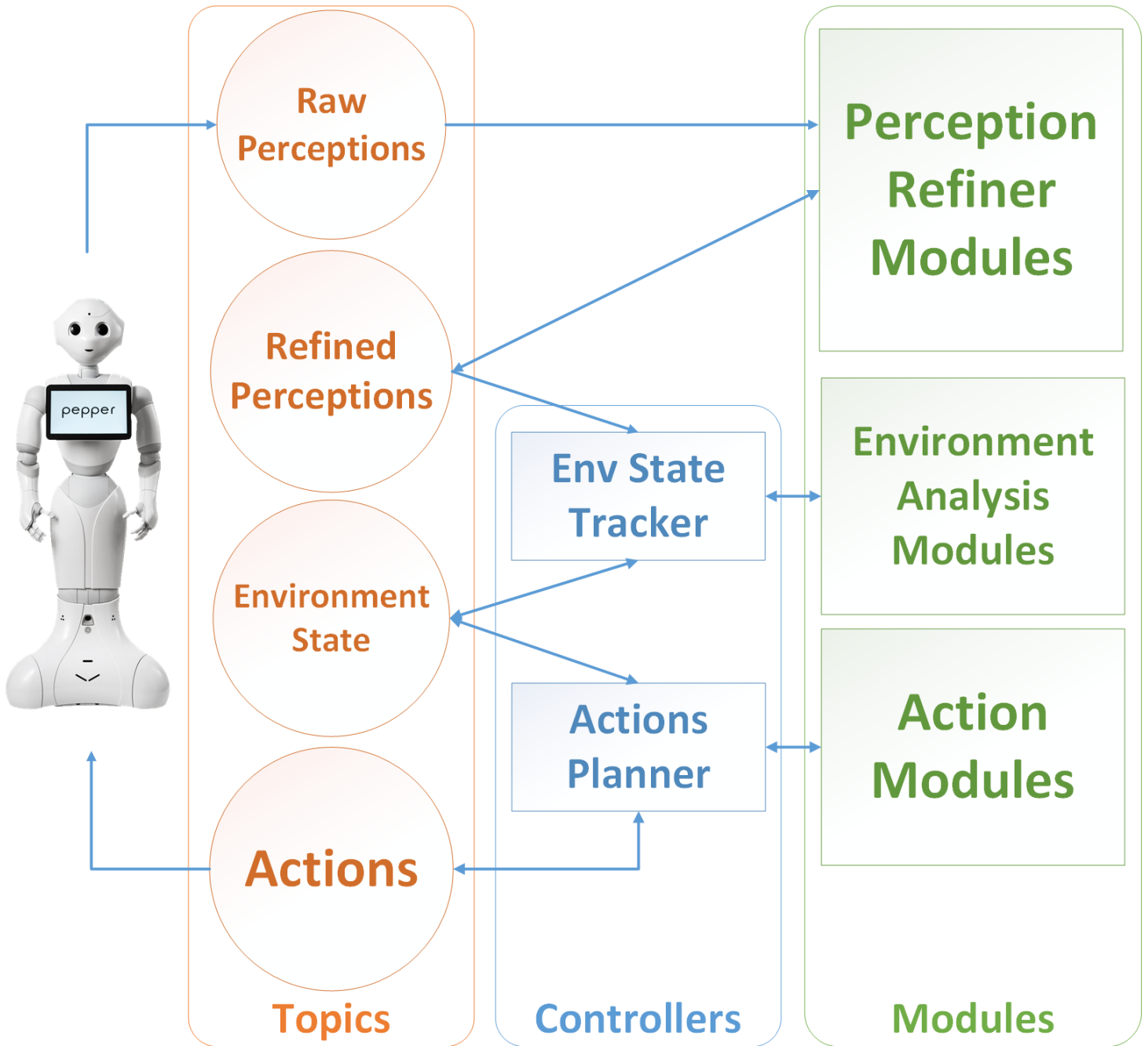


FIGURE 9 – Diagramme de l’architecture FlowAct, une boucle continue de perceptions et d’actions.

Python et C++ qu'en Java, et afin de pouvoir utiliser ROS, dont les langages principaux sont Python et C++, la version 2.5.5 de NAOqi a été choisie pour cette étude.

Ainsi, il est possible de se connecter aux perceptions de la machine à l'aide de la librairie Python SDK, basée sur Python 2. L'extraction et la transmission des perceptions de NAOqi vers ROS sont réalisées grâce à deux paquets disponibles sur GitHub : `nao_interaction`¹ et `ros_ naoqi`². Ces paquets servent de wrappers ROS autour de la couche NAOqi pour permettre l'extraction des données et leur dispatch dans des messages bien formés publiés dans les topics adéquats. Pour faciliter son utilisation, le tout est dockerisé (machine virtuelle locale) dans un OS Ubuntu 16.04, allié avec la version Kinetic de ROS.

Un module personnalisé a été ajouté pour initialiser la langue et la posture du robot, ainsi que pour le redresser périodiquement. Cela est nécessaire car il a été observé que le robot a tendance à s'affaisser après plusieurs déplacements, faisant pointer la caméra vers le sol. Le tout est lancé par un ordinateur distant qui se connecte au broker NAOqi afin de publier l'intégralité des perceptions dans les topics ROS mais aussi d'ouvrir des topics d'actions. Ces topics permettent de déplacer le robot et de le faire parler en publiant directement des messages dans ces topics.

La liste des topics issus du robot ainsi que leur nature est définie dans le tableau suivant :

Topics de perception	Définition
Flux audio	Signal audio brut des 4 microphones
Localisation sonore	Azimut et élévation
Image frontale	Caméra 2D
Distance obstacle	Capteur avant et arrière
Topics d'action	
Déplacement	Mouvements de rotation et translation
Parole	Synthèse vocale vers les 2 haut-parleurs latéraux

TABLE 3 – Tableau des perceptions brutes et des actions.

Ces topics issus du robot peuvent ensuite être exploités et améliorés pour créer de nouveaux états de perceptions, que nous appellerons perceptions raffinées, afin de fournir une représentation précise de l'environnement, nécessaire à la prise de décision.

3.3 RAFFINEMENT DES PERCEPTIONS

3.3.1 Introduction

Pour créer une représentation de la scène la plus précise et interprétable possible pour l'agent, les perceptions brutes ne suffisent plus. Bien que les avancées récentes des modèles LLMs (YE et al. 2024; Z. CHEN et al. 2024; H. LIU et al. 2023) permettent désormais

1. https://github.com/ros-naoqi/nao_interaction/tree/master

2. https://github.com/ros-naoqi/pepper_robot

d'analyser des données brutes telles que les images et les sons, le temps de traitement et les défis énergétiques associés ne les rendent pas toujours optimaux. Et surtout, l'état de développement de ces techniques et leur performance au moment de la réalisation des travaux ne permettaient pas d'envisager leur application immédiate. Ainsi, ces perceptions brutes doivent être transformées en perceptions raffinées grâce à des modules spécifiques qui enrichissent la modélisation de la scène observée en interprétant les données brutes pour délivrer des perceptions contenant un niveau d'information sémantique plus élevé.

Pour comprendre de manière optimale une scène, les modalités audiovisuelles sont essentielles. Nous utiliserons ces deux modalités pour identifier les locuteurs dans la scène, comprendre leur parole, ainsi que pour déterminer la position et les caractéristiques visuelles des personnes présentes. En combinant les informations visuelles et auditives, le système peut fournir une analyse plus complète et précise des interactions et des éléments présents dans l'environnement, indispensable pour la prise de décision.

3.3.2 Parole

3.3.2.1 Détection de l'activité vocale

Le son capté par les microphones est une forme de donnée acoustique qui comprend toutes les vibrations sonores présentes dans l'environnement. Ces vibrations sont converties en signaux électriques par le microphone, puis en données numériques pour une analyse et un traitement ultérieur. Ces données sonores peuvent inclure des voix humaines, des bruits de fond, des sons naturels ou des bruits artificiels. L'utilisation d'un système de détection d'activité vocale (VAD) est cruciale dans ce contexte.

Le VAD permet de distinguer les segments de la signalisation sonore qui contiennent de la parole de ceux qui ne contiennent que du bruit de fond. Cela est essentiel pour plusieurs applications, notamment les communications vocales, les systèmes de reconnaissance vocale et les appareils d'enregistrement audio. En filtrant les sons non pertinents, le VAD améliore la qualité et l'efficacité des systèmes de traitement de la parole, réduit la consommation de bande passante et de stockage et augmente la précision des algorithmes de reconnaissance vocale.

Ainsi, un module de VAD est implémenté dans ROS pour traiter le signal sonore reçu par les microphones de Pepper. Il est extrait du paquet Sparc³. Le fonctionnement est standard, un seuil d'énergie est ajusté dynamiquement en fonction de l'énergie moyenne des trames reçues. Dès que le seuil est dépassé, on enregistre et on coupe l'enregistrement dès que l'on repasse en dessous du seuil. Le segment de parole collecté est alors converti et envoyé au module de reconnaissance de la parole (ASR). En pratique, ce module souffre du bruit interne de Pepper qui augmente le seuil de détection calculé. Ainsi, il est difficile de récupérer des segments audio si la personne chuchote ou ne parle pas suffisamment fort.

L'un des défis rencontrés avec le robot Pepper est l'absence de gestion d'écho. Pour éviter d'ajouter des traitements supplémentaires sur la voix du robot, le VAD est désactivé pendant que le robot parle. La durée de cette désactivation est proportionnelle à la

3. <https://sparc.readthedocs.io/en/latest/runandconf.html>

longueur de la phrase, légèrement supérieure au temps d'élocution de l'agent, et dure généralement entre une et quatre secondes au maximum à la fin.

3.3.2.2 *Reconnaissance de la parole*

La reconnaissance de la parole est une technologie fondamentale dans le domaine de l'interaction humain-robot. Elle permet aux robots de comprendre et d'interpréter les paroles des utilisateurs, rendant ainsi la communication plus naturelle et intuitive. Pour l'implémentation de la reconnaissance de la parole, deux solutions ont été envisagées.

La première solution utilise un service web pour capturer le segment audio et recourt à l'API Google Speech-to-Text⁴. La technologie proposée par Google, connue pour sa rapidité et ses performances, s'adapte par ailleurs assez bien au bruit de fond généré par le moteur des microphones de Pepper. Cependant, cette approche présente des inconvénients, notamment sa dépendance à une connexion internet stable et les problèmes de confidentialité liés à l'envoi de données personnelles, telles que l'identification vocale du locuteur, vers des serveurs externes non propriétaires. Ce souci de confidentialité rend cette solution inutilisable dans le cadre des expérimentations dans un hôpital prévu pour le projet μ DialBot.

Aussi un module alternatif de reconnaissance de la parole a été implémenté. Il s'agit d'un module basé sur ROS, qui récupère le segment audio du VAD dans un topic ROS et l'envoie localement pour traitement au modèle Whisper (RADFORD et al. 2022) d'OpenAI. Ce modèle de Transformers a été pré-entraîné de manière faiblement supervisée sur environ 700 000 heures de données audio étiquetées. Cinq tailles de modèles sont disponibles, chacune offrant un équilibre différent entre le temps d'inférence et la précision. Le modèle "medium" a été choisi, avec ses 769 millions de paramètres, afin de limiter la latence du système. Son temps d'inférence est deux fois plus rapide que celui du modèle "large", ce qui compense la légère baisse de performance.

L'étude de X. CHEN et al. (2024) montre que le petit modèle de Whisper présente un taux d'erreur de reconnaissance (WER) plus élevé que celui de Google Cloud ASR. Cependant, nous avons constaté que les taux d'erreurs respectifs n'ont pas d'impact significatif sur le fonctionnement des modules en aval. Une évaluation plus approfondie de cette question reste toutefois à mener.

3.3.3 *Vision*

3.3.3.1 *Détection de personnes*

Dans le cadre de la conception d'un robot social proactif destiné à interagir avec les personnes dans une salle d'attente, l'intégration d'un module de détection de personnes s'avère essentielle. Surtout, dans le cas de Pepper, celui-ci ne dispose pas nativement de solution effective pour comprendre son environnement afin de s'y mouvoir, telle que la localisation et la cartographie simultanées (Simultaneous Localization and Mapping (SLAM)). Module permettant aux robots de percevoir et de localiser les individus présents dans leur environnement, facilitant ainsi des interactions plus naturelles et ciblées. En

4. <https://cloud.google.com/speech-to-text/docs?hl=fr>

détectant les personnes, le robot peut adapter son comportement et ses réponses en fonction de la présence et des mouvements des utilisateurs, améliorant ainsi l'efficacité et la pertinence de ses interactions.

Au sein du champ des algorithmes de reconnaissance de personnes, l'éventail de solutions proposées est à la fois large et diversifié. Après une évaluation approfondie, notre choix s'est porté sur YOLOv7 (C.-Y. WANG, BOCHKOVSKIY et LIAO 2022), reconnu actuellement comme l'un des modèles de détection d'objets en temps réel les plus efficaces. Plus spécifiquement, nous avons utilisé la variante YOLOv7-pose, qui, en plus de détecter les individus, fournit également les points clés du corps, une fonctionnalité essentielle pour notre analyse. En complément, nous avons intégré l'algorithme DeepSort (WOJKE, BEWLEY et PAULUS 2017) pour faciliter le suivi en temps réel. La combinaison de YOLOv7-pose et DeepSort a permis une détection et un suivi visuels, robustes et efficaces dans notre étude.

L'architecture de YOLOv7, un algorithme de détection d'objets rapide, repose sur les fondations des modèles YOLO précédents, tels que YOLOv4 (BOCHKOVSKIY, C.-Y. WANG et LIAO 2020), Scaled YOLOv4 (C.-Y. WANG, BOCHKOVSKIY et LIAO 2021) et YOLO-R (C.-Y. WANG, YEH et LIAO 2021), et introduit une série d'optimisations et de nouvelles fonctionnalités pour améliorer les performances. Les principaux composants et innovations de l'architecture de YOLOv7 sont les suivants :

- "Efficient Layer Aggregation Network" (ELAN) qui permet de contrôler les chemins de gradient pour un apprentissage efficace dans un réseau plus profond ;
- "Extended Efficient Layer Aggregation Network" (E-ELAN) est le principal bloc de calcul de YOLOv7, il améliore la capacité d'apprentissage du réseau sans perturber la trajectoire originale du gradient grâce à une approche de type expansion, brassage, fusion de la cardinalité ;
- YOLOv7 utilise RepConvN pour éviter les connexions d'identité lors du remplacement des couches convolutives par un résidu ou une concaténation, ce qui permet de résoudre les problèmes de perte de précision constatés dans d'autres architectures ;
- Une tête auxiliaire est implémenté pour aider lors de l'entraînement, ainsi qu'un mécanisme d'attribution d'étiquettes pour une meilleure attribution des étiquettes en tenant compte à la fois des prédictions du réseau et de la vérité de terrain.

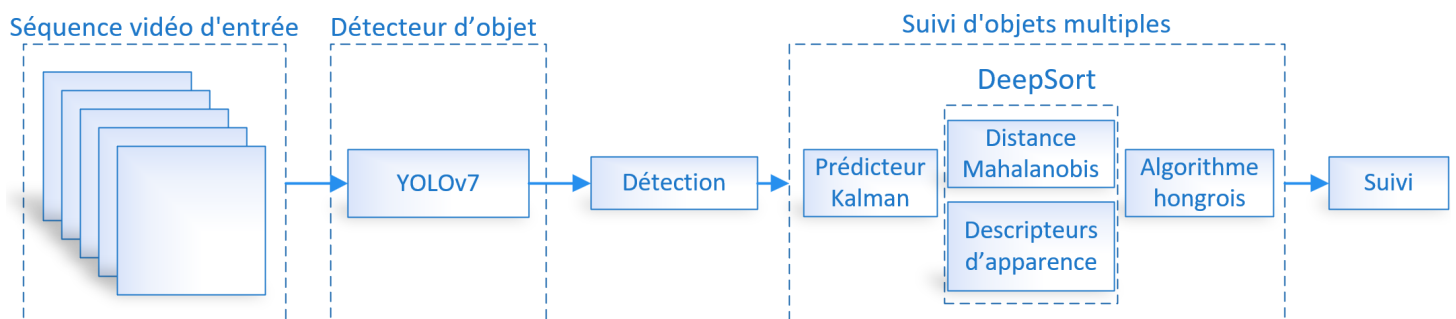


FIGURE 10 – Diagramme en bloc du module de détection de personne.

L'algorithme DeepSort utilise un filtre de Kalman (KALMAN 1960) pour prédire l'emplacement futur des objets détectés sur la base de leurs emplacements antérieurs. Puis, afin d'intégrer les informations relatives au mouvement, la distance de Mahalanobis (MCLACHLAN 1999) est calculée entre les états de Kalman prédits et les mesures nouvellement obtenues. La distance de Mahalanobis permet de tenir compte de l'incertitude de l'estimation de l'état en mesurant le nombre d'écart-types entre la détection et l'emplacement moyen de la piste. Un réseau de neurones profonds pré-entraîné est utilisé pour extraire des vecteurs de caractéristiques à partir des boîtes de délimitation détectées. Ces vecteurs de caractéristiques servent de descripteurs d'apparence des objets. La distance cosinus entre les différents descripteurs d'apparence constitue la deuxième métrique utilisée. L'algorithme hongrois est ensuite appliqué pour associer les emplacements prédits aux nouvelles détections. Cet algorithme résout le problème de l'affectation de manière optimale afin de minimiser le coût de calcul (distance entre les emplacements prédits et les nouvelles détections).

Ainsi, nous avons implémenté un module ROS de détection de personnes (Figure 10) basé sur un modèle YOLOv7 combiné à DeepSort. Ce module permet, à partir d'un flux vidéo, de récupérer les détections d'humains ainsi que leurs identifiants de suivi dans l'image. Il peut être noté que, malgré l'utilisation d'un historique de descripteurs d'apparence, l'algorithme DeepSort a du mal à reconnaître une personne qui sort de l'image pendant une période prolongée ou qui modifie son apparence, par exemple en enlevant un vêtement.

3.3.3.2 *Suivi du regard*

Le module de suivi de regards utilise le paquet ROS de RT Gene (FISCHER, CHANG et DEMIRIS 2018) qui transforme le flux d'images en différents descripteurs du visage : l'orientation du regard, la pose de la tête et la position des points clés du visage. Le modèle en charge du regard est entraîné sur la base de données RT Gene. Dans ces données, l'angle de vue est capturé à l'aide de lunettes de suivi du regard et un réseau sémantique d'incrustation d'images est utilisé pour supprimer les lunettes de suivi des yeux.

L'approche du modèle utilisé exploite les capacités des réseaux convolutifs en cascade multitâches (MTCNN) (K. ZHANG et al. 2016) pour identifier les régions du visage, en capturant spécifiquement les points de repère associés aux yeux, au nez et à la bouche. La taille du segment de visage est ajustée en s'appuyant sur les points de repères identifiés. L'objectif est de réduire l'écart entre ces points de repère et les points de référence standard d'un visage moyen. Ce processus d'alignement utilise l'algorithme itératif accéléré du point le plus proche pour obtenir une représentation normalisée du visage (BESL et MCKAY 1992). Une fois le visage normalisé, des régions rectangulaires de taille fixe sont délimitées autour des points de repère correspondant aux yeux, en les extrayant de l'imagerie faciale normalisée. Dans la phase suivante, pour déterminer l'orientation de la tête, la technique avancée proposée par PATACCHIOLA et CANGELOSI (2017) est utilisée.

3.3.4 *Diarisation multimodale du locuteur*

3.3.4.1 *Introduction*

Dans cette étude, nous présentons un modèle de fusion temporelle audio-visuelle pour la diarisation de locuteurs multi-utilisateurs, visant à répondre à la question "qui parle à un moment donné", que nous avons entièrement proposé et développé.

Dans des domaines tels que la robotique, cette tâche est cruciale pour l'analyse de scènes. Les solutions présentées dans la sous-section 2.4.3 présentent plusieurs désavantages vis-à-vis des scénarios robotiques :

- les modèles sont difficilement transportables et intégrables directement sur le robot;
- ils requièrent plusieurs caméras dans la pièce afin de suivre le locuteur dominant.

Afin de pallier ses limitations, nous proposons une méthode légère basée sur la caméra intégrée à l'agent robotique ainsi que la localisation sonore.

La méthode que nous proposons modélise la fusion des indices audio et visuels par la coïncidence spatiale de la localisation des sources visuelles et sonores (SSL, Sound Source Localization). En combinant ce modèle de concordance avec une formulation bayésienne dynamique, nous pouvons suivre efficacement l'identité du locuteur actif au fil du temps. La SSL offre plusieurs avantages dans les conversations multi-utilisateurs, notamment la possibilité de gérer des segments de parole qui se chevauchent, ce qui élimine le besoin d'un module d'activation vocale. Le modèle génératif sous-jacent, qui ne nécessite pas de phase d'apprentissage, exploite les informations spatiales de SSL et la localisation des visages pour identifier et suivre les locuteurs dominants dans diverses conditions acoustiques. Les paramètres sont estimés en ligne, ce qui garantit de faibles exigences en matière de calcul et des performances robustes. Nous avons validé l'efficacité de notre approche en utilisant deux ensembles de données : un ensemble public et un autre collecté en interne avec le robot humanoïde Pepper.

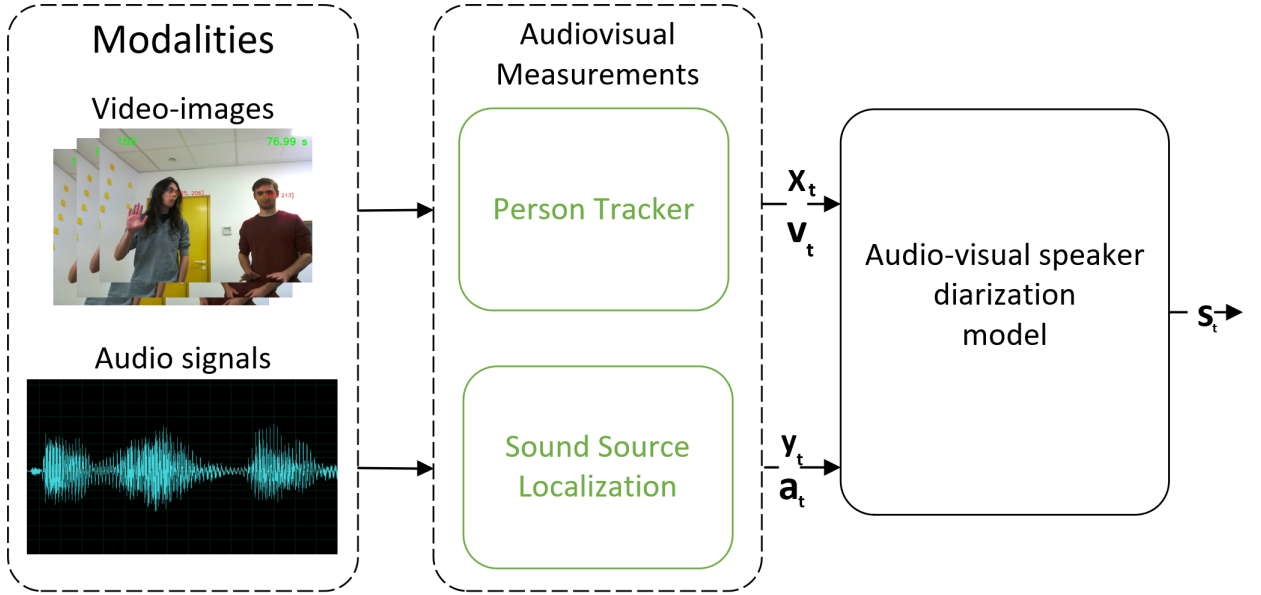


FIGURE 11 – Diagramme en bloc de la méthode de diarisation du locuteur proposée.

3.3.4.2 Méthode proposée

DÉFINITION DU PROBLÈME

Nous représentons t l'indice de pas de temps des images visuelles et audio, synchronisées entre elles. À l'image t , il y a au plus N observations visuelles,

$$\underline{X}_t = (\mathbf{X}_{t1}, \dots, \mathbf{X}_{tn}, \dots, \mathbf{X}_{tN}) \in \mathbb{R}^{2 \times N}$$

, où la variable aléatoire \mathbf{X}_{tn} correspond à l'emplacement de la bouche de la personne n dans l'image t . Un système de suivi de plusieurs personnes fournit alors une série temporelle de N emplacements d'images, à savoir $\underline{X}_{1:t} = \{\underline{X}_1, \dots, \underline{X}_t\}$ et les masques binaires de présence visuelle associés $\underline{V}_{1:t}$, à savoir la variable V_{tn} associée à \mathbf{X}_{tn} de sorte que $V_{tn} = 1$ si la personne n est présente dans l'image à l'instant t et 0 sinon. $N_t = \sum_n V_{tn}$ représente le nombre de personnes observées à l'image t . En pratique, lorsque $V_{tn} = 0$, $\mathbf{X}_{tn} = \mathbf{X}_{t-i,n}$ avec $t-i$ la période la plus récente où $V_{t-i,n} = 1$. Nous considérons également un module [SSL](#) qui fournit l'azimut et l'élévation de la source sonore dominante à chaque trame audio t . L'emplacement de la source sonore peut ensuite être cartographié sur le plan de l'image, de sorte qu'une paire d'observations azimut-élévation est transformée en un emplacement d'image modélisé par une variable aléatoire $\underline{Y}_t = (\mathbf{Y}_{t1}, \dots, \mathbf{Y}_{tk}, \dots, \mathbf{Y}_{tK}) \in \mathbb{R}^{2 \times K}$ avec K observations audiovisuelles pour une image visuelle à t et $\underline{Y}_{1:t} = \{\underline{Y}_1, \dots, \underline{Y}_t\}$ sa série temporelle. À ces observations audiovisuelles, nous associons un masque binaire d'activité vocale $\mathbf{A}_{1:t} = \{\mathbf{A}_1, \dots, \mathbf{A}_t\}$, tel que $A_t = 1$ s'il existe une source audio active à l'image t ou 0 dans le cas contraire.

L'objectif est de suivre le(s) locuteur(s) dominant(s) à l'instant t en associant la responsabilité audio au fil du temps à l'activité audio (le cas échéant) avec l'une des personnes suivies. Les sources audio en dehors des images ne sont pas prises en compte. On parle

également de diarisation audiovisuelle du locuteur, abordée ci-dessous dans le cadre des modèles graphiques temporels.

Nous introduisons une série temporelle de variables latentes discrètes, $\mathcal{S}_{1:t} = \{S_1, \dots, S_t\}$ telle que $S_t = n, n \in 1, 2, \dots, N$ si la personne n est à la fois observée et parle à l'image t , et $S_t = 0$ si aucune des personnes visibles ne parle à l'image t . Notez que $S_t = 0$ représente deux cas différents : premièrement, il y a au moins une source sonore active à t ($A_t = 1$), mais sa localisation ne peut pas être associée à l'une des personnes visibles et elle peut être interprétée comme du bruit ; deuxièmement, il n'y a pas de source sonore active à t , $A_t = 0$.

Nous utiliserons également une autre variable latente $Z_{t1:K} = Z_{t1}, \dots, Z_{tK}$ avec $Z_{tk} = n$ qui représente l'attribution de la source sonore k à l'identité visuelle n . $Z_{tk} = 0$ signifie que la source k n'est attribuée à aucune personne dans l'image.

Variable	Description	Domaine
t	Indice de pas de temps des images visuelles et audio	\mathbb{N}
\underline{X}_t	Emplacement des bouches dans l'image t	$\mathbb{R}^{2 \times N}$
\mathbf{X}_{tn}	Emplacement de la bouche de la personne n dans l'image t	\mathbb{R}^2
$\underline{X}_{1:t}$	Séries temporelles des emplacements d'images jusqu'à t	$(\mathbb{R}^{2 \times N})^t$
$\underline{V}_{1:t}$	Masques binaires de présence visuelle jusqu'à t	$\{0, 1\}^{N \times t}$
V_{tn}	Présence visuelle de la personne n dans l'image t	$\{0, 1\}$
N_t	Nombre de personnes observées dans l'image t	\mathbb{N}
\underline{Y}_t	Emplacement des sources sonores dominantes dans l'image t	$\mathbb{R}^{2 \times K}$
$\underline{Y}_{1:t}$	Séries temporelles des emplacements des sources sonores jusqu'à t	$(\mathbb{R}^{2 \times K})^t$
$\mathbf{A}_{1:t}$	Masques binaires d'activité vocale jusqu'à t	$\{0, 1\}^t$
A_t	Présence d'une source audio active dans l'image t	$\{0, 1\}$
$\mathcal{S}_{1:t}$	Séries temporelles des variables latentes du locuteur dominant jusqu'à t	$\{0, 1, \dots, N\}^t$
S_t	Variable latente représentant la personne parlant à l'image t	$\{0, 1, \dots, N\}$
$Z_{t1:K}$	Attribution des sources sonores aux identités visuelles	$\{0, 1, \dots, N\}^K$
Z_{tk}	Attribution de la source sonore k à l'identité visuelle n	$\{0, 1, \dots, N\}$

TABLE 4 – Tableau résumé des variables utilisées dans le modèle.

MODÈLE DE DIARISATION DU LOCUTEUR

Le problème de la diarisation temporelle du locuteur peut être formulé comme la recherche d'une solution maximale a posteriori (MAP), c'est-à-dire la recherche de la configuration la plus probable de l'état latent S_t qui maximise la distribution de pro-

babilité a posteriori suivante. Également appelée distribution de filtrage, elle peut être exprimée de la manière suivante :

$$\hat{s}_t = \arg \max_{s_t} P(S_t = n | \underline{y}_{1:t}, \underline{x}_{1:t}, \underline{v}_{1:t}, \mathbf{a}_{1:t}) \quad (9)$$

Selon la formule de Bayes, la probabilité a posteriori (9) peut être écrite comme suit :

$$P(S_t = s_t | \mathbf{u}_{1:t}) = \frac{P(\mathbf{u}_t | S_t = s_t, \mathbf{u}_{1:t-1}) P(S_t = s_t | \mathbf{u}_{1:t-1})}{P(\mathbf{u}_t | \mathbf{u}_{1:t-1})} \quad (10)$$

puis pour $\mathbf{u}_t = (\underline{x}_t, \mathbf{v}_t, \underline{y}_t, \mathbf{a}_t)$ développée de la manière suivante :

$$P(S_t = s_t | \mathbf{u}_{1:t}) = \frac{P(\mathbf{u}_t | S_t = s_t) G_{s_t}}{\sum_{j=0}^N P(\mathbf{u}_t | S_t = j) G_j} \quad (11)$$

Avec $G_n = \sum_{i=0}^N P(S_t = n | S_{t-1} = i) P(S_{t-1} = i | \mathbf{u}_{1:t-1})$

L'évaluation de (11) est récursive et un nombre raisonnable de personnes suivies simultanément doit être pris en compte (5-8) pour que le calcul reste traitable. Le calcul de cette équation nécessite la vraisemblance observée $P(\mathbf{u}_t | S_t = s_t)$ et les probabilités de transition $P(S_t = j | S_{t-1} = i)$ expliquées dans les deux sous-sections suivantes.

MODÈLE D'OBSERVATION AUDIO-VISUELLE EM

La principale caractéristique du modèle proposé est sa capacité à associer de manière robuste la SSL à l'instant t à une personne. La maximisation de l'espérance pour le modèle de mélange gaussien déduit la probabilité postérieure qu'une personne prononce un discours à partir d'observations audio et visuelles qui sont cartographiées dans le même espace mathématique. Nous distinguons deux cas. Le premier, s'il n'y a pas d'activité audio au moment t ($A_t = 0$), la probabilité a posteriori peut être évaluée à l'aide de la formule suivante, où c est un petit scalaire positif, par exemple $c = 0,2$:

$$P(S_t = n | \mathbf{y}_t, \underline{x}_t, \mathbf{v}_t, A_t = 0; c) = \begin{cases} c/N_t & \text{if } 1 \leq n \leq N \\ 1 - c & \text{if } n = 0. \end{cases} \quad (12)$$

Si une source sonore est active à l'instant t , ($A_t = 1$), nous lui attribuons une identité visuelle n telle que $Z_{tk} = n$ joue le rôle d'une variable d'affectation dans un modèle de mélange. Son emplacement \mathbf{y}_{tk} est supposé provenir du mélange gaussien/uniforme suivant :

$$P(\mathbf{y}_{tk} | \underline{x}_t, \mathbf{v}_t, A_t = 1; \theta_t) = \sum_{n=1}^N p_{tn} v_{tn} \mathcal{N}(\mathbf{y}_{tk} | \mathbf{x}_{tn}, \underline{\Sigma}_{tn}) + p_{t0} \mathcal{U}(\beta) \quad (13)$$

Où $\theta_t = (\{p_{tn}\}_{n=0}^N, \{\underline{\Sigma}_{tn}\}_{n=0}^N, \beta)$ désigne l'ensemble des paramètres du modèle, à savoir l'a priori, $\sum_{n=1}^N v_{tn} p_{tn} + p_{t0} = 1$, les matrices de covariance 2×2 $\underline{\Sigma}_{tn}$, et le paramètre β qui caractérise la composante du mélange, à savoir une distribution uniforme.

L'ensemble des paramètres θ_t peut être estimé via l'algorithme Expectation Maximization (EM) pour les mélanges gaussiens.

L'algorithme commence par l'étape E qui évalue les probabilités postérieures $r_{tkn} = P(Z_{tk} = n | \mathbf{y}_{tk}, \underline{\mathbf{x}}_t, \mathbf{v}_t, A_t = 1)$ en utilisant les valeurs des paramètres actuels θ_t , Z_t est notre variable d'affectation, $Z_{tk} = n$ signifie que \mathbf{y}_{tk} est généré par le composant n . Nous calculons d'abord $r_{tkn} \forall n, 1 \leq n \leq N$ qui correspond au fait qu'une source sonore est associée à une personne visible :

$$r_{tkn} = \frac{p_{tn} v_{tn} \mathcal{N}(\mathbf{y}_{tk} | \mathbf{x}_{tn}, \underline{\Sigma}_{tn})}{\sum_{i=1}^N p_{ti} v_{ti} \mathcal{N}(\mathbf{y}_{tk} | \mathbf{x}_{ti}, \underline{\Sigma}_{ti}) + p_{t0} \mathcal{U}(\beta)} \quad \forall n, 1 \leq n \leq N \quad (14)$$

On peut aussi écrire la probabilité qu'une source sonore ne soit pas associée à une personne visible $n = 0$, soit parce qu'elle correspond à un son émis par une personne non visible, soit parce qu'elle est émise par un autre type de source, c'est-à-dire la postériorité de la composante uniforme du mélange :

$$r_{tk0} = \frac{p_{t0} \mathcal{U}(\beta)}{\sum_{i=1}^N p_{ti} v_{ti} \mathcal{N}(\mathbf{y}_{tk} | \mathbf{x}_{ti}, \underline{\Sigma}_{ti}) + p_{t0} \mathcal{U}(\beta)} \quad (15)$$

L'étape M pour maximisation réestime les paramètres en utilisant les responsabilités actuelles :

$$\underline{\Sigma}_{tn}^{new} = \frac{1}{R_{tn}} \sum_{k=1}^K r_{tkn} (\mathbf{y}_{tk} - \mathbf{x}_{tn})(\mathbf{y}_{tk} - \mathbf{x}_{tn})^T + \varepsilon \mathbf{I} \quad (16)$$

$$p_{tn}^{new} = \frac{R_{tn}}{K} \quad (17)$$

avec $\varepsilon > 0$ est un scalaire agissant comme un paramètre pour éviter les clusters vides, et \mathbf{I} est la matrice identité 2×2 et où nous avons défini :

$$R_{tn} = \sum_{k=1}^K r_{tkn} \quad (18)$$

L'algorithme peut être facilement initialisé en fixant tous les aprioris à $1/N + 1$ et en fixant toutes les variances à un scalaire positif. Comme les moyennes des composants sont fixes, l'algorithme converge en quelques itérations seulement.

Nous avons N faces et K sources, ce qui représente un problème combinatoire à chaque itération. Pour chaque association possible à t , nous devons considérer tous les cas possibles pour l'étape suivante. Ces calculs entre les visages et les sources vont exploser au fil du temps. Pour résoudre ce problème et garder le modèle audiovisuel traitable, au lieu de calculer toutes les combinaisons, nous factorisons toutes les sources en une source dominante $\mathbf{y}_{tn_t^*}$. Il faut d'abord choisir la personne dont la probabilité de parler est la plus élevée, représentée par l'apriori :

$$n_t^* = \max_n p_{tn} \quad (19)$$

Par conséquent, la source moyenne $\mathbf{y}_{tn_t^*}$ est l'emplacement de la source sonore considéré comme le plus probable sur la base de $\mathbf{x}_{tn_t^*}$:

$$\mathbf{y}_{tn_t^*} = \frac{\sum_{k=1}^K r_{tkn_t^*} \mathbf{y}_{tk}}{R_{tn_t^*}} \quad (20)$$

$P(S_t = n | \mathbf{y}_{t:n_t^*}, \underline{\mathbf{x}}_t, \mathbf{v}_t, \mathbf{a}_t)$ peut alors être calculée à partir de $\mathbf{y}_{t:n_t^*}$, $\forall n, 1 \leq n \leq N$:

$$P(S_t = n | \mathbf{u}_{t:n_t^*}) = \frac{\pi_{tn} \nu_{tn} \mathcal{N}(\mathbf{y}_{t:n_t^*} | \mathbf{x}_{tn}, \underline{\Sigma}_{tn})}{\sum_{i=1}^N \pi_{ti} \nu_{ti} \mathcal{N}(\mathbf{y}_{t:n_t^*} | \mathbf{x}_{ti}, \underline{\Sigma}_{ti}) + \pi_{t0} \mathcal{U}(\beta_t)} \quad (21)$$

et pour $n = 0$

$$P(S_t = 0 | \mathbf{u}_{t:n_t^*}) = \frac{\pi_{t0} \mathcal{U}(\beta_t)}{\sum_{i=1}^N \pi_{ti} \nu_{ti} \mathcal{N}(\mathbf{y}_{t:n_t^*} | \mathbf{x}_{ti}, \underline{\Sigma}_{ti}) + \pi_{t0} \mathcal{U}(\beta_t)} \quad (22)$$

avec $\mathbf{u}_{t:n_t^*} = (\mathbf{y}_{t:n_t^*}, \underline{\mathbf{x}}_t, \mathbf{v}_t, \mathbf{a}_t)$.

Enfin, en notant que la vraisemblance des données observées $P(\mathbf{u}_{t:n_t^*})$ ne dépend pas de S_t et en supposant une distribution uniforme sur les antécédents de la personne visible n ($\nu_{tn} = 1$), c'est-à-dire $\pi_{t0} = \pi_{tn} = 1/(N_t + 1)$, nous obtenons le modèle d'observation suivant :

$$\begin{aligned} P(\mathbf{u}_{t:n_t^*} | S_t = n) &= P(S_t = n | \mathbf{u}_{t:n_t^*}) P(\mathbf{u}_{t:n_t^*}) / P(S_t = n) \\ &= P(S_t = n | \mathbf{u}_{t:n_t^*}) P(\mathbf{u}_{t:n_t^*}) / \pi_{tn} \\ &\propto P(S_t = n | \mathbf{u}_{t:n_t^*}). \end{aligned} \quad (23)$$

Cela permet de remplacer la vraisemblance observée (côté gauche de (23)) par la vraisemblance postérieure (côté droit de (23)) dans (11) :

$$P(S_t = s_t | \mathbf{u}_{1:t}) = \frac{P(\mathbf{u}_t | S_t = s_t) G_{s_t}}{\sum_{j=0}^N P(\mathbf{u}_t | S_t = j) G_j} \quad (24)$$

$$\propto \frac{P(S_t = s_t | \mathbf{u}_t) G_{s_t}}{\sum_{j=0}^N P(S_t = j | \mathbf{u}_t) G_j} \quad (25)$$

L'objectif ici étant de récupérer le n avec la plus haute probabilité $P(S_t = n | \underline{\mathbf{y}}_{1:t}, \underline{\mathbf{x}}_{1:t}, \underline{\mathbf{v}}_{1:t}, \mathbf{a}_{1:t})$, passer par une valeur proportionnelle de $P(\mathbf{u}_t | S_t = s_t)$ ne pose pas de problème.

MODÈLE DE TRANSITION D'ÉTAT AUDIOVISUEL

Les probabilités de transition d'état (GEBRU, Silèye BA et al. 2015), $p(S_t = j | S_{t-1} = i)$, fournissent la modalité temporelle pour suivre les tours de parole le long du pas de temps. La probabilité $p(S_t = j | S_{t-1} = i)$ est calculée dans plusieurs cas en fonction de

la présence ou de l'absence de personnes et de leur statut de locuteur (par commodité et sans perte de généralité, nous fixons $v_{t0} = 1$) :

$$P(S_t = j | S_{t-1} = i) = \begin{cases} p_s & \text{if } i = j \text{ and } v_{t-1,i} = v_{ti} = 1 \\ (1 - p_s)/N_t & \text{if } i \neq j \text{ and } v_{t-1,i} = v_{tj} = 1 \\ 0 & \text{if } v_{t-1,i} = v_{t-1,j} = 1 \text{ and } v_{tj} = 0 \\ 1/N_t & \text{if } v_{t-1,i} = 1, v_{ti} = 0 \text{ and } v_{tj} = 1 \\ 1/N & \text{if } v_{t-1,i} = 0 \text{ and } v_{tj} = 0 \end{cases} \quad (26)$$

Le premier cas de l'équation (26) définit la probabilité d'auto-transition, p_s , par exemple $p_s = 0,8$, de la personne i présente à la fois à $t - 1$ et à t . Le deuxième cas définit la probabilité de transition de la personne i présente à $t - 1$ vers une autre personne j présente à t . Le troisième cas interdit simplement les transitions entre une personne i présente à $t - 1$ et une personne j présente à $t - 1$ mais non présente à t . Le quatrième cas représente la probabilité de transition d'une personne i présente à $t - 1$ mais absente à t , vers une personne j présente à t . Le cinquième cas définit la probabilité de transition entre une personne i absente à $t - 1$ et une personne j absente à t . On peut facilement vérifier que $\sum_{j=1}^N p(S_t = j | S_{t-1} = i) = 1$.

ALGORITHME

La mise en œuvre du modèle complet est décrite dans l’algorithme suivant :

Algorithmus 2 : Diarisation sur locuteur

Paramètres : $\beta, \Sigma_t, \varepsilon, c$
Sortie : $s_{1:t}$
Observations : $x_{1:t}, y_{1:t}, v_{1:t}$
// Modèle d’observation audio-visuelle EM

```

1 for  $t = 1$  to  $end$  do
2   if  $A_t = 0$  then
3     | Calculer  $P(S_t|u_t)$  (12)
4   if  $A_t = 1$  then
5     | Calculer  $r_{tk0}$  (15) et  $r_{tkn}$  (14)
6     | Re-estimer les paramètres  $\Sigma_t^{new}$  (16),  $p_t^{new}$  (17)
7     | Calculer  $y_{tn^*}$  (19) et (20)
8     | Calculer  $P(S_t = n|u_{tn^*})$  (21) et (22)
9     // Modèle de transition d’état audio-visuel
10    | Calculer  $P(S_t|S_{t-1})$  (26)
11    // Modèle de graph temporel
12    | Calculer  $P(S_t|u_{1:t})$  (11)
13    | Estimer le speaker dominant (9)

```

3.3.4.3 Expériences

JEU DE DONNÉES

La méthode proposée est évaluée sur deux corpus. Le premier corpus considéré est l’ensemble des données Co-located Audio-Visual streams with 3D tracks (CAV3D) (QIAN et al. 2019) pour le suivi des locuteurs 3D. Il a été acquis à l’aide d’un système de collecte de données comprenant une caméra couleur à objectif unique associée à un réseau de huit microphones disposés en cercle. Ce système capture le son à un taux de haute fidélité de 96 kHz et la vidéo à un taux de 15 images par seconde. Placé sur une table dans une pièce de 4,77 x 5,95 x 4,5 mètres, dont le temps de réverbération est proche de 0,7 seconde, ce dispositif est capable d’enregistrer jusqu’à trois locuteurs en même temps. L’ensemble de données CAV3D est synchronisé, calibré. Il contient 20 séquences dont la durée varie de 15 à 80 s pour un total de 15 mn. Une évaluation est menée sur un premier sous-ensemble appelé Single Object Tracking (SOT), composé de 9 séquences avec un seul locuteur et un deuxième sous-ensemble SOT2 composé de 6 séquences avec un seul locuteur actif et une deuxième personne interférait (i.e. mais ne parlant pas). La sixième séquence de SOT est utilisée pour calibrer les paramètres.



FIGURE 12 – Exemple de diarisation du locuteur sur deux participant du corpus Pepper, le cercle représente le locuteur dominant prédit.

Le second corpus est enregistré par nos soins sur Pepper. Il contient des dialogues entre deux ou trois personnes. Au total, 9 sujets différents ont participé à cette expérience : 2 femmes, 7 hommes. Il a été demandé aux participants de parler un par un et d'essayer d'éviter les chevauchements. Nous avons apporté des variations aux dialogues en demandant aux participants d'entrer et de sortir de la scène de manière aléatoire, de faire face au robot ou de se regarder les uns les autres. Différentes positions dans la pièce ont été utilisées pour obtenir différentes configurations acoustiques. Dans un dialogue à trois personnes, il a été demandé au participant placé au milieu de rester silencieux, afin d'agir comme un distracteur. La durée totale est d'environ onze minutes et le locuteur dominant est soigneusement identifié dans chaque image. Les fenêtres dont la boîte de délimitation n'est pas détectée ou est calculée de manière imprécise sont mises de côté. Il a été observé que le module *SSL* de Pepper connaît parfois des problèmes d'activation et, en l'absence d'étalonnage du module *SSL* de Pepper, les fenêtres contenant de la parole, mais ne détectant pas de *SSL* sont également éliminées en post-traitement. La table 5 récapitule les deux jeux de données.

SPÉCIFICATIONS TECHNIQUES

Pour cette expérience, le modèle de diarisation du locuteur est mis en œuvre pour CAV3D et le corpus Pepper avec quelques différences. Les *SSL* sont extraits à l'aide du paquet ODAS (Open embeddeD Audition System) <http://odas.io>. Il intègre l'algorithme SRP-PHAT-HSDA (GRONDIN et MICHAUD 2018), qui balaie l'espace avec des grilles de résolution grossière et fine pour réduire le nombre de consultations de mémoire, en uti-

Caractéristique	CAV _{3D}	LIA Pepper
Nombre de séquences	20	9
Durée totale	15 minutes	11 minutes
Nombre de sujets différents	2	9
Nombre maximal de locuteurs par scène	1	2
Nombre maximal de personnes par scène	2	3

TABLE 5 – Comparaison des datasets CAV_{3D} et LIA Pepper.

lisant un modèle de directivité du microphone pour diminuer le nombre de directions à scanner et ignorer les paires de microphones non significatives.

L’interpolation est effectuée à partir des coordonnées 3D, données par *SSL*, pour la localisation de la bouche du locuteur en utilisant le sous-ensemble *SOT*. Le module *SSL* de Pepper récupère la direction de la source émettrice (angles d’azimut et d’élévation) à partir des différences de temps d’arrivée, mesurées sur les différentes paires de microphones. Les angles fournis par le moteur de localisation de la source sonore correspondent à la position réelle de la source avec une précision moyenne de 10 degrés. Le passage de l’angle à la localisation dans le plan de l’image se fait par interpolation. Nous enregistrons la *SSL* à partir d’un haut-parleur à différentes positions dans l’image et effectuons une régression pour obtenir un angle de correspondance avec la position de l’image. Les sources sonores situées en dehors de l’image sont filtrées. Nous calibrons et affinons les paramètres de la configuration micro en utilisant la première séquence *SOT* pour *CAV_{3D}* et une séquence d’entraînement pour le corpus Pepper en préparation de la phase de test.

Ainsi, nous avons fixé les paramètres suivants 6 pour les deux expériences :

Paramètres	CAV _{3D}	Corpus Pepper
$\underline{\Sigma}$	Diag[300, 800]	Diag[300, 500]
β	10^7	300000
ϵ	100	200
c	0.2	
p_s	0.8	

TABLE 6 – Tableau résumé des paramètres pour CAV_{3D} et le corpus Pepper.

RÉSULTATS

La performance de la diarisation est évaluée par le taux d’erreur de diarisation (Diarization Error Rate (*DER*)), dans le tableau 7. Il comprend trois termes : Détection manquante (Missing Detection (*MS*)), Fausse alarme (False Alarm (*FA*)) et Erreur de locuteur (Speaker Error (*SPKE*)). Plus il est faible, meilleure est la performance.

Expériences	MS	FA	SPKE	DER
SOT (pas de séquence 6)	3.29	6.63	/	9.92
SOT (pas de séquence 6) oracle	1.24	0	/	1.24
SOT2	6.22	8.02	0	14.2
SOT2 oracle	1.14	0	0	1.14
Corpus Pepper	5.35	0.1	13.82	19.27

TABLE 7 – Tableau des performances (%) de notre modèle pour différents ensembles d’expériences, A_t est défini avec l’oracle VAD dérivé des étiquettes de diarisation, ou avec la présence ou l’absence de SSL.

Pour évaluer les conditions acoustiques, nous étudions l’ensemble de données CAV₃D avec et sans oracle VAD. L’utilisation d’un oracle réduit considérablement le DER car il fournit des informations précieuses pour réduire le nombre de FA. Les résultats sur CAV₃D sont prometteurs, le modèle ne perd que 4,28% de DER entre SOT et SOT2 sans VAD oracle. Avec un DER de 19,27% sur le corpus Pepper, nous pouvons supposer que notre modèle remplit son objectif de diarisation dans un cas robotique standard. Cette méthode montre des résultats intéressants sur SOT2 avec un SPKE de 0%. La prédiction ne correspond à la bonne personne que lorsqu’elle détecte un locuteur. Mais elle diminue considérablement sur l’ensemble de données Pepper. Cela provient de scénarios plus complexes et est également lié à la qualité des microphones du Pepper.



FIGURE 13 – Exemples de diarisation du locuteur sur trois participants, la vérité de base est représentée par un carré, la prédiction par un cercle, $y_{t_n^*}$ est le point rouge et les autres points verts plus petits sont les localisations sonores. Notre modèle fait la bonne prédiction sur les images 1-5, mais échoue sur les images 6-9 à prédire le locuteur précédent.

Ces résultats peuvent être comparés à ceux de l'état de l'art (State Of The Art (SOTA)) de la diarisation audio-visuelle du locuteur : WST (CHUNG, B.-J. LEE et I. HAN 2019) et à la diarisation du locuteur uniquement audio SOTA VBx (LANDINI et al. 2020) sur le corpus AMI (CARLETTA et al. 2005). Le corpus AMI est une collection de réunions qui présente des similitudes avec le corpus Pepper. WTS donne 21,3% et 21,1% de DER sur les deux sous-ensembles AMI : ES, IS et VBx 38,65% pour l'ensemble du corpus AMI. Ces deux résultats sont calculés sans VAD oracle. Nous obtenons des résultats similaires en tenant compte des variations de DER entre les ensembles de données. La complexité théorique de l'algorithme est de $O(n^2)$. Le modèle d'observation audio-visuel représente 99% du temps d'exécution. Pour 10 sources sonores détectées, avec 5 personnes considérées, le temps d'exécution est de 0,0631 secondes, sur un temps d'exécution total de 0,0635 secondes.

3.3.4.4 Conclusion

Nous avons proposé dans cette section un modèle de différenciation temporelle du locuteur basé sur des concepts mathématiques et algorithmiques couplés à deux types de perception, le SSL et l'image plane. Ce modèle donne de bons résultats avec une capacité d'adaptation à différentes conditions acoustiques sans phase d'entraînement. Ainsi, cette méthode de diarisation n'est pas biaisée par un ensemble de données d'entraînement particulier, et elle est donc applicable à un grand nombre de scénarios pratiques d'interaction humain-robot. La fonction VAD et la robustesse sont assurées par la composante uniforme du mélange, qui recueille les emplacements des sources sonores éloignées des composantes gaussiennes, qui sont centrées autour des visages. Le modèle est relativement léger au niveau calculatoire, ce qui en fait un modèle facilement incorporable pour des solutions temps-réel.

Cependant, pour obtenir la fusion audio et image, une interpolation doit être effectuée entre la SSL et le plan d'image pour chaque configuration micro. Parvenir à éliminer cette dépendance est un défi pour les travaux futurs visant à rendre encore plus transférable ce modèle d'un système à un autre.

3.3.5 Analyse de l'engagement en HRI

3.3.5.1 Introduction

L'analyse de l'engagement est une information intéressante dans une analyse de scène. Il existe deux définitions de l'engagement dans la littérature. INOUE et al. (2018) définissent l'engagement comme un état qui suppose que l'on est soit engagé, soit non engagé. L'engagement peut également être vu comme « *Le processus par lequel les interactants initient, maintiennent et mettent fin à leurs connexions perçues les uns avec les autres au cours d'une interaction.* » (SIDNER, C. LEE et LESH 2003). On retrouve l'analyse de l'engagement dans de nombreux contextes tels que les musées (SALAM et CHETOUANI 2015), les jeux (KLOTZ et al. 2011), le milieu hospitalier (SIDNER, CANDACE L. et DZIKOVSKA, MYROSLAVA 2005), l'éducation (PAPADOPOULOS et al. 2016), ou encore la réception (BAEK, CHOI et KWAK 2014).

L'analyse d'un faible taux d'engagement pourrait aider un agent à prendre des décisions lors d'une interaction en choisissant une réponse susceptible d'accrocher l'utilisateur au dialogue. Cela pourrait donner des indications cruciales pour permettre un comportement proactif.

Dans cette étude, nous cherchons à modéliser l'engagement dans un scénario d'interaction humain-robot. Pour cela, nous étudions la possibilité de rapprocher deux jeux de données d'engagement :

- DAISEE (GUPTA et al. 2022), ensemble de données de classification vidéo multi-label pour reconnaître les états affectifs de l'utilisateur tels que l'ennui, la confusion, l'engagement et la frustration dans la nature et
- UE-HRI (BEN-YOUSSEF et al. 2017), un ensemble de données d'interactions spontanées entre un robot et des humains.

3.3.5.2 Présentation des jeux de données utilisés

Le jeu de données DAiSEE est le premier ensemble de données pour la classification vidéo multi-labels. Il comprend 9068 extraits vidéo capturés auprès de 112 utilisateurs. Ces vidéos identifient l'état de l'utilisateur en termes d'ennui, de confusion, d'engagement et de frustration. L'ensemble des données comprend quatre niveaux d'étiquettes (très faible, faible, élevé et très élevé) pour chacun des états affectifs. Les étiquettes ont été annotées par la foule et mises en corrélation avec une annotation standard dorée établie à l'aide d'une équipe de psychologues experts. La durée des vidéos est d'environ 10 secondes et elles sont enregistrées à une fréquence de 30 images par seconde. Les données sont réparties en trois ensembles distincts, à savoir l'ensemble d'entraînement, l'ensemble de validation et l'ensemble de test. La répartition des classes est peu équilibrée, comme le confirme le tableau 8.

	0	1	2	3
Boredom	3822	2850	1923	330
Engagement	61	455	4422	3987
Confusion	5951	2133	741	100
Frustration	6887	1613	338	87

TABLE 8 – Tableau de la répartition des classes par niveau d'expression.

Le jeu de données UE-HRI regroupe des interactions spontanées entre un robot et des êtres humains, parmi lesquelles 54 interactions (d'une durée chacune de 4 à 15 minutes) sont librement accessibles pour le téléchargement et l'utilisation. Les participants ont été filmés lors de conversations spontanées avec le robot Pepper. Les échanges débutaient automatiquement lorsque le robot détectait la présence d'un participant et la conversation était enregistrée si le participant acceptait les conditions (c'est-à-dire, d'être enregistré). Pepper était placé dans un espace public où les participants avaient la liberté d'initier et de conclure l'interaction à leur gré. Les labels liés à l'engagement correspondent à une baisse d'engagement à un instant t , avec 1 si on observe une baisse d'engagement dans les 2 dernières secondes. Le jeu de données est très déséquilibré avec seulement 2% d'étiquettes baisse d'engagement sur l'entièreté du jeu de données en prenant un pas de 0.5 seconde. Des exemples de deux jeux de données sont proposés dans la figure 14. Les trois images supérieures correspondent à un score d'engagement dans l'ordre de 1,2 et 3. Les images du bas correspondent au label 0 du SED formulé dans la sous-section suivante 3.3.5.3.

3.3.5.3 Formulation du problème

Afin de construire un modèle d'engagement utilisable en HRI, le modèle doit être suffisamment rapide en temps d'inférence pour fonctionner en temps réel. Le premier modèle développé répond à un problème de classification multiclassés, il estime le niveau d'engagement (entre 0 et 3) d'une personne à l'instant t , basé sur le comportement observé sur l'intervalle :

$$[t - \tau, t]$$



FIGURE 14 – Exemples de l’ensemble de données Daisee, en haut, et UE-HRI, en bas.

où τ représente les dernières secondes prises en compte. Pour le second, l’objectif est de détecter en temps réel la baisse de l’engagement de l’utilisateur. Autrement dit, étant donné une fenêtre d’observation de durée τ , qui peut contenir des tours de parole de l’utilisateur et du robot, nous cherchons à déterminer si cette séquence manifeste un Signe de Diminution de l’Engagement (*SED*) de la part de l’utilisateur. Pour la tâche de classification binaire ainsi définie, nous considérons toutes les séquences de longueur τ dans le jeu de données UE-HRI. Nous attribuons l’étiquette 1 (resp. 0) si un *SED* est présent (resp. absent) dans les dernières η secondes de la séquence. Il est à noter que le choix des valeurs de τ et η dépend de l’application et de la conception du problème. Dans ce travail, nous fixons τ et η à 5 et 2 secondes respectivement, ce qui correspond, pour un pas de temps fixé à 0,5 seconde, à 10 et 4 intervalles de temps respectivement.

3.3.5.4 Caractéristiques utilisées

Les caractéristiques audio sont extraites à l’aide d’OpenSmile (Florian EYBEN, WÖLLMER et SCHULLER 2010), tandis que les caractéristiques vidéo sont extraites avec OpenFace (BALTRUŠAITIS et al. 2018). Une fenêtre glissante de 0,5 seconde est utilisée pour calculer la moyenne et la variance des caractéristiques. Les caractéristiques d’OpenFace utilisées comprennent l’angle du regard, la pose et la rotation de la tête, ainsi que 17 unités d’action pour un total de 44 caractéristiques visuelles. Quant aux caractéristiques audio, il s’agit de la liste des caractéristiques eGeMAPSv02 d’OpenSmile décrites par F. EYBEN et al. (2016), représentant 25 caractéristiques.

3.3.5.5 Métriques

Pour évaluer les résultats, nous calculons le score F1 ainsi que le score de rappel, utiles pour traiter les jeux de données mal répartis.

Le score F1 quantifie l’équilibre entre la précision (proportion de cas positifs prédits qui sont correctement identifiés comme positifs) et le rappel, fournissant une mesure plus

robuste lorsqu'il s'agit d'ensembles de données déséquilibrés (SOKOLOVA et LAPALME 2009). Il est la moyenne harmonique de la précision et du recall, ce qui se traduit par l'équation suivante :

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

où :

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

True Positive (TP) représente le nombre de vrais positifs, False Positive (FP) représente le nombre de faux positifs, et False Negative (FN) représente le nombre de faux négatifs. Le score F1 varie entre 0 et 1, où 1 indique une performance parfaite du modèle.

3.3.5.6 Expériences et résultats

Le jeu de données déséquilibré de UE-HRI rend l'apprentissage complexe. Ainsi, nous entraînons d'abord un modèle sur DAISEE afin de travailler en apprentissage par transfert. Au regard de la taille des données, le modèle utilisé est un GRU (CHO et al. 2014) d'une couche avec 32 unités cachées, appris sur 100 epochs avec un taux d'apprentissage de $1e^{-4}$. Nous obtenons un score F1 de 0.65 sur le set de test. L'ajout d'une couche d'attention fut testé sans augmentation significative du score F1.

Pour la classification binaire du SED nous utilisons la même architecture de modèle que pour DAISEE. Nous évaluons en plus du score F1 le rappel (ou sensibilité), qui est la proportion des items pertinents proposés parmi l'ensemble des items pertinents. Nous évaluons trois approches : l'approche utilisée pour DAISEE, avec une couche d'attention et avec les valeurs d'engagement de DAISEE. Les résultats sont donnés dans le tableau 9.

Expériences	F1 apprentissage	Rappel apprentissage	F1 test	Rappel test
Modèle 1	0.9908	0.7201	0.9835	0.0291
Modèle 2 (attention)	0.9995	0.9923	0.9739	0.0291
Modèle 3 (engagement)	0.9943	0.8473	0.9756	0.0291

TABLE 9 – Tableau des résultats des différentes approches pour $\tau = 5s$, timestep = 0.5s, $\eta = 2s$.

On observe que le modèle n'arrive pas à généraliser correctement : les SED du set de test ne sont pas détectés alors que ceux du set d'entraînement le sont. Cela est vraisemblablement imputable à la taille du jeu de données qui ne fournit pas assez d'exemples variés du SED. L'ajout des prédictions du modèle entraîné sur DAISEE n'améliore pas les résultats. Les quelques exemples (figure 15) de l'évolution de l'engagement prédit par le modèle entraîné sur DAISEE et la vérité terrain de la baisse d'attention dans le set de donnée UE-HRI ne montrent pas de signe de corrélation convaincant. En effet, si un SED est détecté lors d'une interaction, le niveau d'engagement devrait descendre, ce qui n'est pas le cas dans les exemples.

3.3.5.7 Conclusion

La prédiction d'engagement dans un contexte d'interaction humain-robot, telle que nous l'avons proposée dans cette section, ne fonctionne pas suffisamment bien pour une utilisation dans une configuration réelle de système multimodale. Et ce, malgré la combinaison d'un jeu de données relativement conséquent de l'engagement d'une personne dans un contexte d'apprentissage en ligne et un jeu de données d'interaction spontanée humain-robot. La principale explication provient du manque de données pour le second corpus, couplé à une trop grande différence de nature entre le niveau d'engagement et le [SED](#).

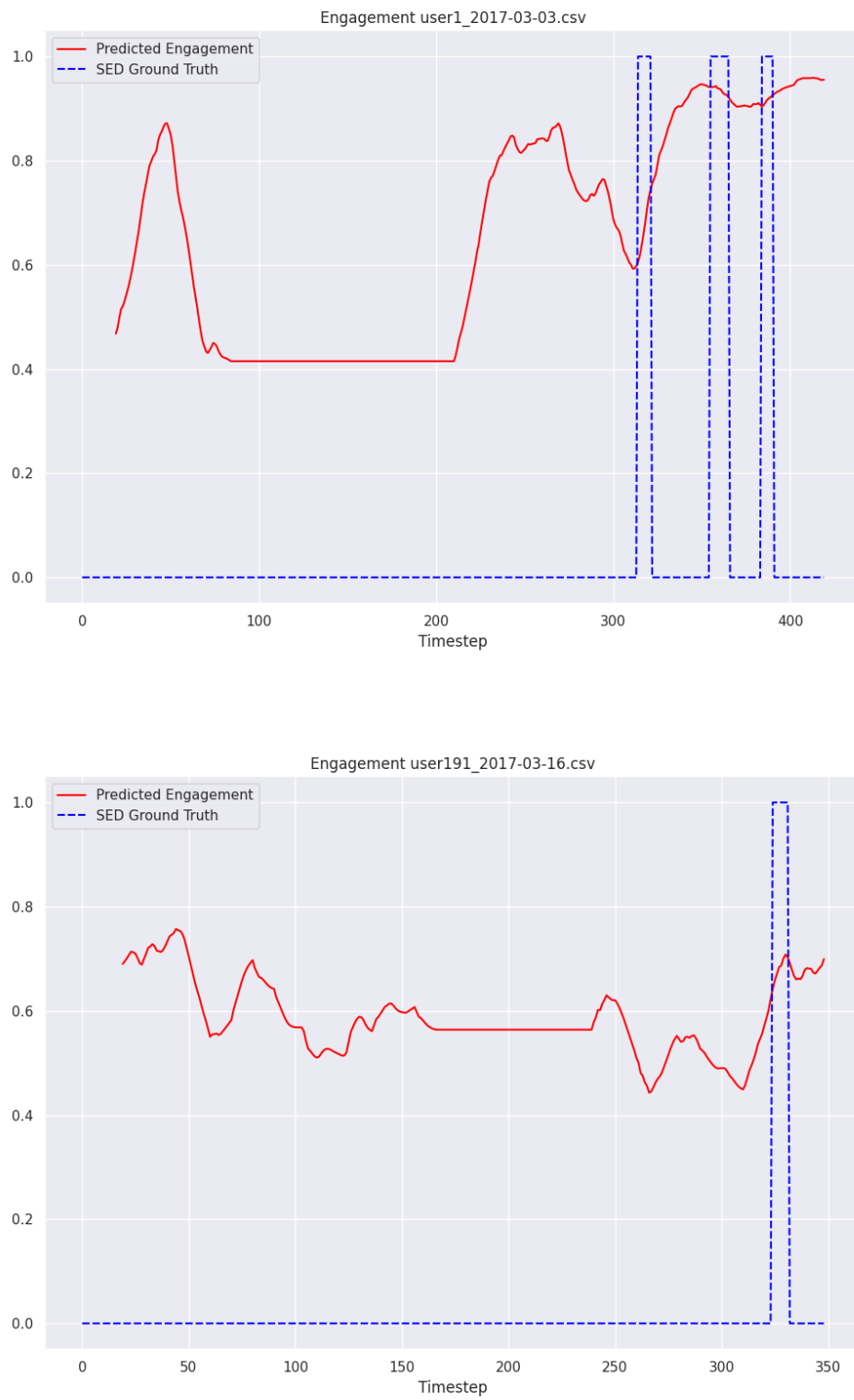


FIGURE 15 – Exemples d’interactions avec un SED détecté (pointillé bleu) et la prédiction d’engagement du modèle DAISEE (ligne rouge).

3.3.6 *Analyse de l'acceptation de l'engagement*

3.3.6.1 *Introduction*

Comprendre le comportement humain dans des contextes sociaux fournit des informations cruciales pour des interactions efficaces. Notamment lorsque les individus envisagent de communiquer, ils évaluent rapidement la probabilité d'un engagement en se basant sur les signaux sociaux et l'activité observée du potentiel partenaire. Nous proposons de nommer ce processus cognitif l'**IAB**, ou croyance d'acceptation de l'interaction.

Le concept d'**IAB** trouve son application dans divers scénarios de robotique sociale, y compris les tâches de service, les approches proactives et les méthodes réactives. Dans cette sous-partie, nous présentons une définition complète de la croyance d'acceptation de l'interaction et proposons une méthodologie pour sa modélisation réaliste dans des scénarios du monde réel.

Notre approche vise à améliorer les capacités des robots sociaux à déduire efficacement les préférences humaines et à s'y adapter, ce qui doit amener à des interactions humain-robot plus efficaces. En menant des évaluations expérimentales, nous établissons la faisabilité du développement d'un modèle qui capture et représente la croyance d'acceptation de l'interaction dans un contexte social spécifique, l'interaction dans une salle d'attente d'un hôpital.

3.3.6.2 *Définition*

La recherche sur l'interaction sociale humain-robot vise à modéliser le concept d'intelligence sociale par le biais d'un dispositif robotique. Reproduire le comportement humain dans un environnement social est une tâche difficile. Un système d'interaction humain-robot (**HRI**) devrait être capable d'interagir naturellement avec une personne tout en étant capable de répondre aux stimuli de son environnement. Ainsi, l'action de l'agent est réalisée par l'analyse des signaux émis par les personnes à sa portée (espace social), au moyen de perceptions multimodales. Les processus et capacités impliqués dans la perception, l'interprétation et la compréhension des informations sociales sont nommés reconnaissance sociale (**WYER et CARLSTON 1979**) et son importance a été soulignée par **SANDINI et al. (2018)** dans le contexte de la **HRI**.

Lorsqu'il s'agit d'initier une interaction avec quelqu'un, le cerveau humain évalue rapidement les chances de succès de sa tentative. Un inconnu est moins susceptible d'accepter l'interaction, tout comme une personne déjà engagée dans une autre interaction ou occupée à une tâche sans rapport avec l'agent. Cette analyse de la disponibilité de la personne aide à décider de s'engager ou d'attendre une meilleure occasion. Le comportement de la personne joue un rôle important dans cette évaluation, car nous pouvons estimer qu'environ 60 à 65% de toutes les communications ou interactions interpersonnelles sont constituées de comportements non verbaux (**BURGOON 1994**). L'activité de l'utilisateur peut également donner des indications sur la disponibilité de la personne. En outre, l'activité et le comportement non verbal peuvent être par définition observés visuellement. Contrairement à d'autres caractéristiques, telles que l'état mental de l'individu ou sa proximité personnelle avec l'agent, ces caractéristiques n'ont pas de mani-

festations externes à l'individu ou sont trop complexes pour être déduites uniquement d'une image.

Dans cette section, nous proposons les bases du concept d'Interaction Acceptance Belief : **Interaction Acceptance Belief (IAB)** - Un concept qui répond à la question :

"Quelles sont les chances que mon interaction soit acceptée par l'utilisateur ciblé?"

Il représente une mesure de l'incertitude concernant le niveau d'acceptation d'une interaction potentielle entre un utilisateur ciblé et l'agent. Son expression peut se manifester à travers le comportement passif ou actif de l'utilisateur vis-à-vis de l'agent.

L'IAB ouvre de nombreuses possibilités de recherche sur l'interaction sociale humain-robot, dans le but de modéliser le concept d'intelligence sociale par le biais d'un robot. Il apporte des informations cruciales pour les agents robotiques ayant besoin d'initier une interaction de manière proactive (GROSINGER 2022). La nécessité d'un raisonnement transparent (BREMNER et al. 2019b) pour les actions des robots basées sur des perceptions et des croyances est un sujet important dans l'interaction humain-robot et dans cette logique l'IAB permet d'explicitier le choix d'un robot engageant une personne plutôt qu'une autre dans un scénario d'interaction proactive.

3.3.6.3 Modélisation

Nous visons à effectuer une prédiction de l'IAB dans un environnement propice dans le contexte de l'HRI proactive, comme l'est notre objectif opérationnel la salle d'attente d'hôpital. Les patients attendent leur tour de manière plutôt statique, ils peuvent être debout ou assis sur une chaise, utiliser leur téléphone ou simplement écouter de la musique. Les personnes levant la main demandent une interaction; cette interaction n'a presque aucune chance d'être rejetée, sauf dans de rares cas d'appels erronés. Une personne souriant à l'agent peut conduire à une interaction proactive réussie. Les deux situations aboutissent à un résultat similaire, où l'interaction est acceptée. Cependant, le niveau d'incertitude pour la prise de décision varie entre elles.

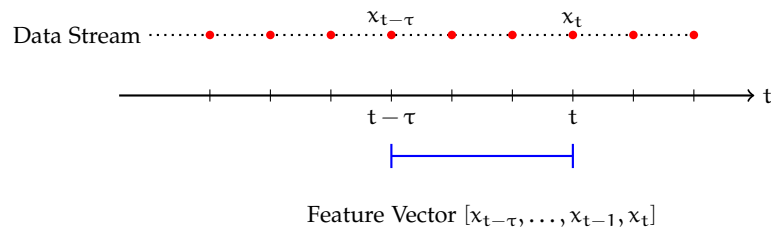


FIGURE 16 – Diagramme de la modélisation de l'IAB.

La prédiction de la valeur de l'IAB est modélisée de la manière suivante (figure 16). Une fenêtre d'observation est définie comme un intervalle de temps $[t - \tau, t]$, qui capture les dernières τ secondes du comportement de l'utilisateur jusqu'au temps t . Cette fenêtre sert de base pour générer un vecteur de caractéristiques $[x_{t-\tau}, \dots, x_{t-1}, x_t]$, qui inclut les images dans l'intervalle. Ce vecteur de caractéristiques est utilisé comme entrée pour

le classificateur. La sortie du classificateur attribue une étiquette à chaque fenêtre d'observation, indiquant le degré auquel l'utilisateur peut être engagé pendant cette période.

Au pas de temps t , nous construisons un modèle qui classe le comportement de l'utilisateur observé dans l'intervalle $[t - \tau, t]$ comme "peut accepter une interaction" ou "peut ne pas accepter une interaction". Laissons $X = [x_1, x_2, \dots, x_T]$ représenter la séquence de vecteurs de caractéristiques du comportement utilisateur multimodal, et laissons $Y = [y_1, y_2, \dots, y_T]$ désigner la séquence correspondante d'étiquettes de sorties binaires, avec $y_t = C([x_{t-\tau}, \dots, x_{t-1}, x_t])$, C étant le classificateur.

3.3.6.4 Pré-traitement

Pour prédire les niveaux de l'IAB, une approche multi-niveaux est utilisée, impliquant l'extraction de quatre niveaux de perceptions différents. Ces perceptions comprennent :

CARACTÉRISTIQUES DE LA TÊTE Le premier niveau de perception consiste à analyser le regard, la position de la tête et l'unité d'action de l'individu. Les unités d'action font référence aux mouvements des muscles faciaux ou aux configurations associées à des expressions faciales spécifiques (TADAS BALTRUŠAITIS et ROBINSON 2015). Ces informations permettent de savoir où la personne regarde et quelle est l'orientation de sa tête. Un outil possible pour déduire l'orientation du regard et la pose de la tête est OpenFace⁵, une boîte à outils d'analyse du comportement facial en libre accès (BALTRUŠAITIS et al. 2018).

POINTS CLÉS DU CORPS Le deuxième niveau de perception étudie l'impact des points clés du corps dans l'image et la manière dont ils sont liés à des poses corporelles particulières, afin de permettre une compréhension plus approfondie de la relation entre le langage corporel et les niveaux d'IAB. Ces caractéristiques sont extraites par image avec le modèle YoloV7 (C.-Y. WANG, BOCHKOVSKIY et LIAO 2022).

CARACTÉRISTIQUES D'ACTION Le troisième niveau de perception consiste à extraire les caractéristiques liées aux actions de la personne. Il peut s'agir d'analyser les gestes, les mouvements du corps ou d'autres indices comportementaux qui fournissent des informations sur l'activité en cours de la personne, y compris l'engagement ou l'interaction. Un modèle I3D (CARREIRA et ZISSERMAN 2018), pré-entraîné sur l'ensemble de données Charades (SIGURDSSON et al. 2016), extrait des caractéristiques d'action par le biais d'une fenêtre coulissante de 64 images sur des images d'utilisateurs délimitées par leurs boîtes de délimitation. Le vecteur de caractéristiques extrait a une longueur de 1024.

ÉMOTIONS Le quatrième niveau de perception se concentre sur l'extraction des états émotionnels. Il s'agit d'analyser les expressions faciales de la personne afin de déterminer leur valence émotionnelle et leur niveau d'excitation. En outre, les cinq expressions primaires, à savoir Neutre, Heureux, Triste, Surprise et Peur, peuvent être détectées. La détection des émotions est assurée par une implémentation de (TOISOUL et al. 2021) et l'extraction est effectuée par image.

Pour obtenir des vecteurs de caractéristiques synchronisés, nous utilisons l'intégration temporelle (également connue sous le nom de regroupement temporel) en appli-

5. <https://github.com/TadasBaltrusaitis/OpenFace>

quant une fenêtre d'intégration commune à tous les flux de caractéristiques. Le processus d'intégration implique l'application d'une fonction d'intégration spécifique, telle que la moyenne et la variance, sur des fenêtres d'intégration glissantes d'une longueur de L secondes. Dans cette étude, des fonctions d'intégration basées sur les statistiques sont utilisées, en particulier, les fonctions de moyenne et de variance. La longueur de la fenêtre d'intégration L est fixée à 500 ms et il n'y a pas de chevauchement entre les fenêtres d'intégration.

3.3.6.5 *Jeu de données*

Le jeu de données utilisé pour ce projet a été collecté pour créer une simulation de patients dans une salle d'attente. Pour ce faire, nous avons fait appel à des acteurs non professionnels qui ont reçu pour instruction de jouer des comportements tout en étant enregistrés. Un acteur dédié, représentant le point de vue de l'agent, se déplaçait dans la pièce avec une caméra positionnée au niveau du torse. Parfois, il était demandé aux acteurs de changer leurs scénarios de comportement pour introduire de la variété dans la scène et être représentatifs de situations réelles. La durée des scènes varie entre 30 s et 2 min, pour une durée totale d'environ une heure.

La liste suivante présente une liste exhaustive des comportements potentiels des acteurs considérés dans cette étude :

- Engager des conversations avec des individus assis à côté d'eux.
- Utilisation active des téléphones mobiles, comme jouer à des jeux ou naviguer sur Internet.
- Montrer un comportement passif, où ils restent inactifs et semblent attendre sans engagement ou activité spécifique.
- Montrer des signes d'intérêt et d'attention envers l'agent.
- Chercher l'attention ou l'assistance de l'agent, demander des informations, des conseils ou du soutien.

Un groupe diversifié d'acteurs, comprenant 12 individus de genres différents, mais tous âgés de [20, 30], a été sélectionné pour jouer le rôle des patients dans l'étude. Pour simplifier le processus et garantir la reproductibilité, les expériences sont enregistrées avec la caméra frontale d'un Apple iPhone 13 à 30 images par seconde. Le niveau de l'IAB de chaque acteur de la vidéo a été méticuleusement étiqueté sur une échelle de 1 à 5 par deux annotateurs, un doctorant et un alternant en master d'informatique. La consigne donnée aux annotateurs de données est la suivante : un segment de temps avec un minimum de 2 secondes correspondant à un comportement est annoté, lorsque le comportement change un nouveau segment est annoté. Plus il est proche de 1, plus il est peu probable qu'il accepte l'interaction. Inversement, plus il est proche de cinq, plus il est susceptible d'accepter une interaction. Un exemple est également fourni aux annotateurs pour chaque niveau. Après l'étiquetage, le coefficient Kappa de Cohen entre les annotateurs est de 0,88. Des exemples de comportements tirés des ensembles de données sont illustrés dans la figure 17, pour laquelle les labels du niveau de l'IAB sont, de haut en bas, de gauche à droite : 3, 5, 2, 1 pour les deux acteurs, 4 et 2.

Le jeu de données présente un déséquilibre significatif des classes (table 10), avec les étiquettes de l'IAB 4 et 5 représentant environ 14% de la distribution globale des classes.



FIGURE 17 – Exemples de comportements issus du jeu de données.

Niveau de l'IAB	1	2	3	4	5
Durée (s)	1532	1295	1521	391	322
Pourcentage	30.26%	25.60%	30.05%	7.72%	6.36%

TABLE 10 – Tableau de la répartition des valeur de l'IAB dans le jeu de données.

Cependant, cela découle naturellement des acteurs jouant leurs scénarios et choisissant le moment d'exprimer leur intérêt pour une interaction.

3.3.6.6 Expérience

Le modèle utilisé dans notre expérience est basé sur l'architecture "many to one" Gate Recurrent Unit (GRU) (figure 18), telle qu'introduite par CHO et al. (2014). Les GRU sont un type de réseau de neurones récurrents (RNN) avec des mécanismes de portes supplémentaires qui aident à contrôler le flux d'informations au sein du réseau. Ces mécanismes de contrôle permettent aux GRU de mieux gérer le flux d'informations dans le temps et d'atténuer le problème du gradient de disparition, ce qui leur permet de capturer les dépendances à long terme plus efficacement que les RNN traditionnels. L'architecture consiste en une seule couche GRU avec 128 dimensions cachées, qui agissent comme des cellules de mémoire pour stocker et propager l'information dans le temps.

Dans le contexte d'un ensemble de données de petite taille, le modèle peut facilement devenir trop sensible à des exemples d'apprentissage spécifiques, ce qui entraîne une instabilité pendant l'apprentissage. Des variations mineures ou des valeurs aberrantes dans les données limitées peuvent avoir un impact significatif sur la représentation apprise par le modèle. Le modèle peut également éprouver des difficultés à saisir des modèles com-

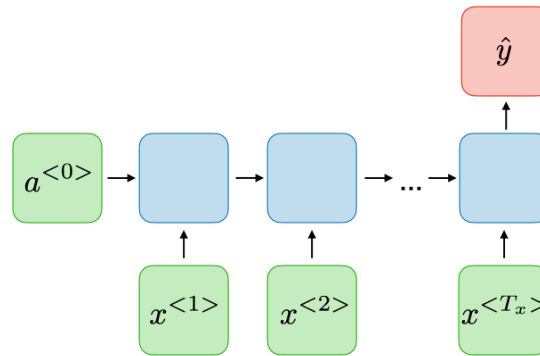


FIGURE 18 – Diagramme des cellules GRU en configuration 'many-to-one'.

plexes : les GRU sont conçus pour apprendre des modèles et des relations complexes au sein des données. Cependant, avec trop peu de données, le modèle peut ne pas avoir suffisamment d'exemples diversifiés pour saisir toute la complexité du problème, ce qui conduit à des performances sous-optimales. Pour relever ces défis, nous utilisons une technique de sur-échantillonnage dans l'ensemble d'apprentissage afin d'augmenter les données et d'obtenir une représentation plus équilibrée des étiquettes. Une validation croisée six fois est employée pour exploiter les données disponibles, le meilleur modèle étant conservé. Les modèles sont ensuite évalués sur un ensemble de tests de 10 minutes.

3.3.6.7 Métriques

L'étude initiale se concentre sur l'évaluation de la performance d'un réseau neuronal récurrent classique à travers différentes modalités. Pour évaluer les résultats, nous calculons l'aire sous la courbe caractéristique de fonctionnement du récepteur (AUROC) et le score F1.

L'AUROC sert d'indicateur pour évaluer les performances des modèles de classification binaire (Fawcett 2006). Il fournit une évaluation complète de la capacité du modèle à différencier les classes en fonction de diverses valeurs de seuil. En tenant compte de l'équilibre entre le rappel (proportion de cas positifs correctement identifiés) et la spécificité (proportion de cas négatifs correctement identifiés), l'AUROC quantifie le pouvoir discriminant global du modèle. L'équation de l'AUROC est donnée par :

$$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(x)) dx$$

où le TPR (Taux de Vrais Positifs) est tracé en fonction du FPR (Taux de Faux Positifs) à différents niveaux de seuil. Mathématiquement, le TPR est défini comme $\frac{\text{TP}}{\text{TP} + \text{FN}}$ et le FPR est défini comme $\frac{\text{FP}}{\text{FP} + \text{TN}}$, où TP, FP, TN et FN représentent respectivement les vrais positifs, les faux positifs, les vrais négatifs et les faux négatifs. L'intégrale calcule effectivement la surface sous la courbe ROC, fournissant une valeur scalaire qui résume la capacité du modèle à discriminer entre les classes positive et négative.

3.3.6.8 Résultats

Nous avons effectué des tests avec différentes caractéristiques et leurs combinaisons. Les résultats obtenus pour chaque caractéristique individuelle et les 3 meilleures combinaisons de caractéristiques sont présentés dans le tableau 11.

Caractéristiques/fenêtre d'observation		F1	AUROC
Tête	6s	0.59	0.75
Position du corps	3s	0.52	0.72
Action	5s	0.37	0.61
Émotion	4s	0.36	0.5
Tête+Position du corps	4s	0.61	0.8
Tête+Position du corps+Émotion	2s	0.63	0.79
Tête+Position du corps+Action+Émotion	6s	0.57	0.75

TABLE 11 – Tableau des résultats des scores F1 et AUROC sur différentes combinaisons dans la fenêtre de temps optimale.

Les caractéristiques d'action et d'émotion sont dans l'ensemble peu performantes. Les points clés du corps fournissent quelques indices au modèle pour évaluer l'IAB d'une situation. Avec un AUROC de 0,7, le modèle de pose du corps a une capacité moyenne à différencier les cas entre les cas positifs et négatifs. Les caractéristiques de la tête et de la pose du corps apparaissent comme des éléments communs parmi les trois premières combinaisons de caractéristiques. Cette observation souligne l'importance constante de ces indices perceptifs dans la performance globale du modèle prédictif. La présence des caractéristiques de la tête et de la pose du corps dans les combinaisons les plus performantes renforce leur rôle fondamental dans la capture d'informations pertinentes et dans le discernement des modèles d'acceptation de l'interaction. Les caractéristiques d'action n'ont pas d'influence significative sur la performance prédictive ou ne contribuent pas à l'amélioration de la prédiction générale lorsqu'elles sont combinées à d'autres caractéristiques.

Une observation remarquable découlant de nos expériences est qu'une proportion notable des modèles les plus performants présente des longueurs de fenêtre égales ou supérieures à 4. Cette constatation suggère que la tâche de prédiction de l'IAB bénéficie de la collecte d'un plus grand nombre d'observations jusqu'à ce qu'un certain seuil soit atteint. Au fur et à mesure que le contexte temporel augmente grâce à l'utilisation de fenêtres plus longues, les modèles prédictifs semblent gagner un avantage pour discerner avec précision la dynamique de l'IAB.

La combinaison Tête+Taille+Émotion donne le meilleur score F1, obtenu avec une fenêtre de seulement 2 secondes. Ce résultat intrigant suggère que cette combinaison particulière a atténué le sur-ajustement jusqu'à un certain point tout en facilitant la reconnaissance de scènes spécifiques où le comportement de l'acteur a évolué pour montrer de l'intérêt pour l'agent et a appelé activement à l'interaction. L'utilisation d'indices émotionnels en conjonction avec les caractéristiques de la tête et de la pose du corps semble améliorer légèrement la capacité du modèle à capturer les changements compor-

tements notables et les expressions d'intérêt, contribuant ainsi à la prédiction réussie de l'acceptation de l'interaction dans de tels cas.

3.3.6.9 Conclusion

Cette partie présente et définit l'Interaction Acceptance Belief (IAB) dans le domaine de l'interaction humain-robot. Nous avons introduit une nouvelle approche de modélisation pour la prédiction de l'IAB et l'avons validée de manière approfondie en utilisant un ensemble de données réelles comprenant divers scénarios.

Les résultats obtenus sont prometteurs et encourageants. Notamment, notre recherche souligne le rôle critique joué par la pose de la tête et du corps, en particulier lorsqu'ils sont combinés, dans l'obtention de prédictions précises de l'IAB. L'intégration de ces indices perceptifs contribue de manière significative à la capacité du modèle à discerner et à comprendre la dynamique de l'acceptation de l'interaction dans divers contextes du monde réel.

Cette étude s'est concentrée sur la faisabilité de la prédiction de l'IAB, justifiant ainsi l'utilisation d'un modèle simple de GRU. Les résultats obtenus montrent que ce modèle offre une base solide pour la prédiction initiale de l'IAB. Toutefois, pour améliorer la précision et la performance des prédictions, l'intégration de modèles plus complexes, tels que les modèles de langage étendus, représente une voie prometteuse pour les recherches futures.

Ce signal social est une composante essentielle de la proactivité, car il permet aux systèmes robotiques de mieux anticiper et de répondre aux besoins des utilisateurs. L'intégration de l'IAB dans les modèles d'interaction humain-robot contribue également à l'évolution des robots vers des agents plus empathiques et socialement intelligents. En détectant et en interprétant correctement les signaux d'acceptation ou de refus d'interaction, les robots peuvent éviter les comportements intrusifs ou inappropriés, favorisant ainsi des interactions plus harmonieuses et respectueuses.

3.3.7 Résumé intermédiaire

Pepper est équipé pour saisir un éventail de données multimodales émanant de son environnement, que ses capteurs convertissent en informations exploitables. Ces informations sont acheminées via des topics vers une suite de modules dédiés à l'épuration et à l'enrichissement de ces signaux primaires, incluant notamment les données sonores et visuelles. Un écosystème de modules a été développé pour assurer des fonctions avancées telles que le suivi de personnes, la détection de points clés corporels, l'identification d'individus, la reconnaissance vocale, la diarisation de locuteurs, la reconnaissance automatique de la parole, la traduction, la détection du regard et l'évaluation de l'engagement dans l'interaction. C'est également dans cette partie que nous développons un outil essentiel pour faire interagir l'agent de manière proactive, l'IAB. Ces modules renforcent la représentation de l'environnement, mais se nourrissent aussi parfois entre eux. Par exemple, des modules comme la diarisation du locuteur tirent parti des données déjà traitées, telles que l'identification constante et précise des personnes dans l'environnement, pour élaborer de nouvelles interprétations analytiques.

L'ensemble des modules sont implémentés via ROS ou en tant que services web, ces derniers étant privilégiés lorsque l'intégration des topics ROS n'est pas nécessaire, ils sont récapitulés dans la tableaux ci-dessous (12) :

Topics de perception extraites	Définitions
Voix détectée	Détection de l'activité de la voix
Parole transcrite	Reconnaissance automatique de la parole
Boîte englobante, points clés du corps, identifiant visuel	Détection de personne
Azimut/élévation du regard, orientation de la tête	Suivi du regard
Identifiant locuteur dominant	Diarisation du locuteur
Croyance d'acceptation d'interaction par identifiant	Analyse de l'engagement

TABLE 12 – Tableau récapitulatif des perceptions extraites.

L'intégration de ces informations raffinées avec les données brutes confère au module de suivi environnemental la capacité de former une représentation dynamique et continuellement mise à jour de l'espace ambiant. Cette représentation est cruciale pour les modules décisionnels qui, s'appuyant sur elle, élaborent des stratégies d'action pour interagir avec des utilisateurs.

3.4 FLOWACT, UNE ARCHITECTURE PROACTIVE D'INTERACTION HUMAIN-ROBOT

3.4.1 Fonctionnement général

Dans la sous-section suivante, nous décrivons les spécificités de FlowAct, un système architecturé (figure 19) pour agir comme un conduit perpétuel pour les entrées sensorielles et les actions qui en résultent.

La structure globale suit une architecture standard de HRI basé sur la boucle perception, représentation et action (MOULIN-FRIER et al. 2018; ROSSI et al. 2022). L'agent en contact avec le monde exécute une boucle infinie avec comme entrée les senseurs et comme sortie les effecteurs. L'architecture cognitive passe par trois étapes, la **perception** du monde via les senseurs, la **représentation** de la scène interne à l'agent et la prise de **décision**.

FlowAct sépare les perceptions en deux niveaux, les "perceptions brutes" et les "perceptions raffinées" et ajoute un mécanisme modulaire de sélection des actions. Le concept de perception brute, un terme que nous employons pour désigner les données sensorielles brutes de l'agent, reflétant le quintet des sens humains, notamment l'auditif et le visuel, ainsi que les perceptions directement dérivées de ces sens, tel que la profondeur. Ces perceptions élémentaires sont interceptées par les "modules de raffinement de la perception", un ensemble dans lequel les données brutes sont distillées, produisant une couche d'interprétation améliorée de l'environnement que nous appelons les perceptions raffinées.

Le contrôleur de suivi de l'état de l'environnement opère de manière continue, intégrant à la fois les perceptions brutes et raffinées pour modéliser une représentation dynamique de l'environnement. Responsable de la gestion de la mémoire, il peut se connecter

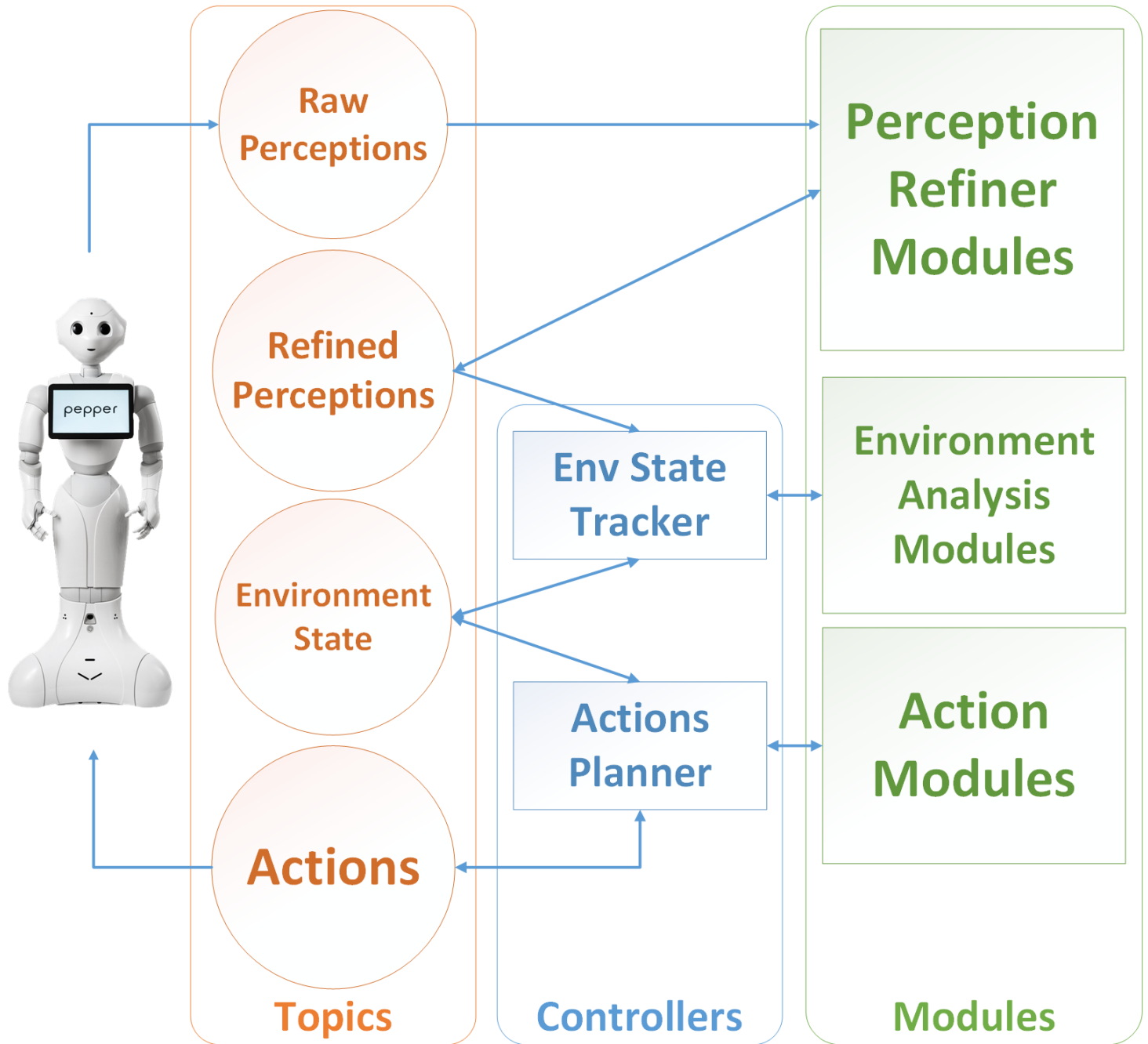


FIGURE 19 – Diagramme de l’architecture FlowAct, une boucle continue de perception et d’action.

à une base de données. Ce contrôleur fait appel à divers modules environnementaux, tels que la ré-identification des individus et l'assignation des perceptions, afin d'actualiser la scène en tenant compte de son historique et de son contexte.

L'état de l'environnement est présenté au 'planificateur d'actions', le centre stratégique de FlowAct. Ce module est chargé de l'analyse des données environnementales et de la planification des stratégies d'action en informant les modules d'action de manière asynchrone. Il possède la capacité unique de réfléchir sur les actions futures tout en exécutant les actions actuelles. Les modules d'action, agissant comme les effecteurs du système, transmettent des directives comportementales précises à l'agent robotique, réalisant ainsi la transition de l'état de l'environnement à l'action au sein de cette boucle autonome.

3.4.2 Perception

3.4.2.1 Le Raffinement de Perception

L'approche FlowAct vise à transformer les perceptions primaires et brutes de l'agent en perceptions améliorées ou raffinées. Les perceptions brutes sont comparables aux cinq sens humains, qui relèvent de l'extéroceptif, ainsi qu'à la perception, consciente ou non, de la position des différentes parties du corps sans recours à la vision, ce que l'on nomme la proprioception. Les perceptions raffinées, quant à elles, résultent du traitement opéré par les 'modules de raffinement de la perception' (figure 20).

Ces modules traitent et améliorent les données initiales, produisant un ensemble de perceptions plus avancé. Les perceptions raffinées seront combinées avec les perceptions brutes, puis envoyées au suivi de l'état de l'environnement, qui construira une représentation de la scène à partir de ces données. Les modules de perceptions raffinées travaillent de manière asynchrone et sans interruption. Sauf cas particuliers, comme par exemple l'arrêt du détecteur de voix lorsque l'agent parle, peut permettre d'éviter d'envoyer des segments audio inutiles pour la compréhension de la scène à la reconnaissance automatique de la parole.

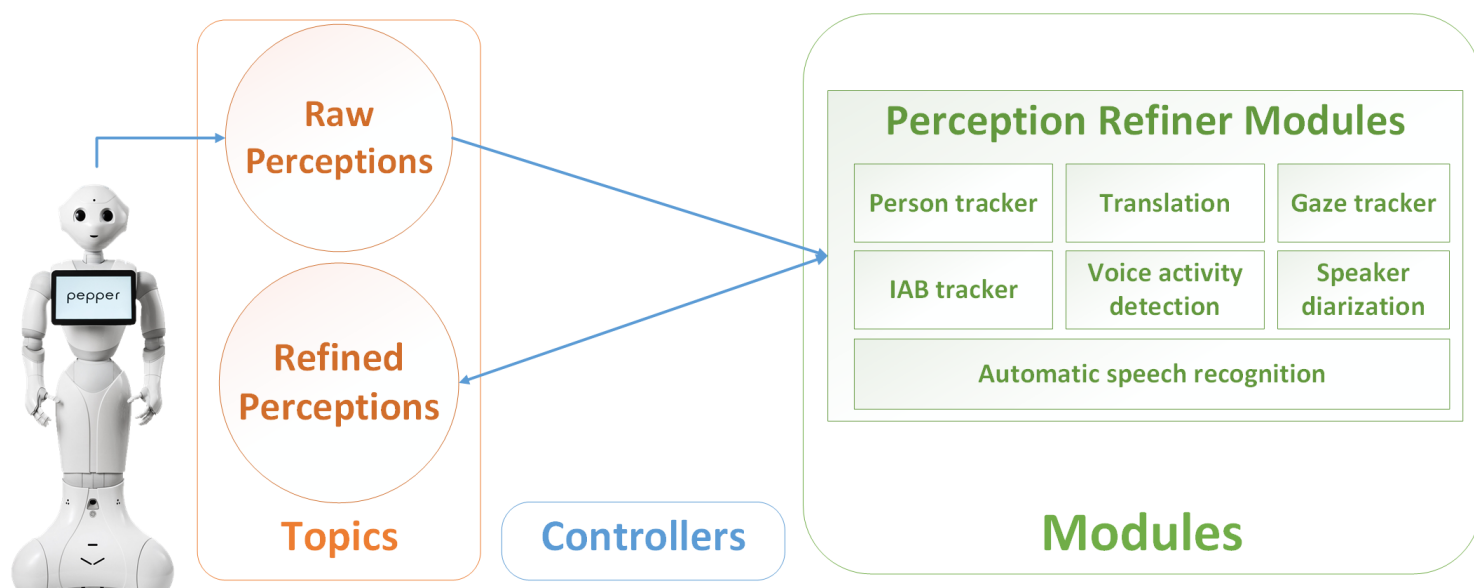


FIGURE 20 – Diagramme de la couche de perception de FlowAct.

3.4.2.2 Les modules de perceptions raffinées

Les modules de perceptions sont vitaux pour obtenir une première vue de l'environnement. Ils récupèrent les informations données par les senseurs de l'agent afin de les retravailler et de livrer une couche de perception supplémentaire appelée les perceptions raffinées.

Pour notre implémentation de l'architecture FlowAct, les modules utilisés sont ceux décrits dans la section précédente et résumés ci-dessous :

- **Détection d'activité vocale** : basée sur un seuil adaptatif de l'énergie du signal sonore.
- **Suivi de personne** : combine la localisation de la personne et les caractéristiques des points clés du corps à partir d'une implémentation de YoloV7 (C.-Y. WANG, BOCHKOVSKIY ET LIAO 2022) avec un algorithme Deepsort (WOJKE, BEWLEY ET PAULUS 2017) pour suivre l'identité dans l'image.
- **Suivi du regard** : utilise le package ROS de RT Gene (FISCHER, CHANG ET DEMIRIS 2018) qui transforme le flux d'images en divers descripteurs faciaux, orientation du regard, posture de la tête et position des points clés du visage.
- **Diarisation des locuteurs** : repose sur un modèle de fusion audio-visuelle temporelle pour la diarisation multi-utilisateurs des intervenants (DHAUSSY, JABAIAH, LEFÈVRE ET HORAUD 2023), la méthode identifie les locuteurs dominants et les suit dans le temps en mesurant la coïncidence spatiale entre les emplacements sonores et la présence visuelle.
- **Croyance d'acceptation de l'interaction** : infère le niveau de l'IAB (DHAUSSY, JABAIAH ET LEFÈVRE 2023) qui répond généralement à la question « Quelles sont les chances que mon interaction soit acceptée par l'utilisateur ciblé? », principalement basée sur le regard de l'utilisateur.

- **Reconnaissance de la parole** : fonctionnalités utilisant des appels API qui transmettent les segments de parole détectés par le module VAD au service de reconnaissance vocale de Google pour la transcription automatique⁶. En alternative, la version large de Whisper (RADFORD et al. 2022) est employée lors des expériences à l'hôpital Broca afin de respecter les normes de confidentialité.
- **Traduction automatiques** : dans le cas où le système de chatbot est basé sur une langue différente, le webhook de traduction DeepL⁷ est utilisé.

Ainsi, l'ensemble de ces modules fournissent une perception améliorée de la scène, mais désordonnée. Ce flux de perceptions raffinées doit être coordonné et assemblé dans la couche de représentation, responsable de la modélisation de l'environnement.

3.4.3 La modélisation de l'environnement

Le suiveur (*tracker*, ou "Contrôleur de Suivi") de l'État de l'Environnement (figure 21) est responsable de construire une représentation précise de la scène au fil du temps. Il collecte les perceptions raffinées, maintient un historique, orchestre le flux des perceptions et offre une représentation complète de l'environnement. C'est dans cette collecte asynchrone de perceptions que le Contrôleur de Suivi de l'État de l'Environnement aligne et organise les perceptions pour publier de manière synchrone une représentation de l'environnement. Le suiveur est aussi chargé de la mémoire associative (PRESCOTT et al. 2019b), un concept définissant la capacité de relier deux éléments initialement sans rapport, tels qu'un nom et un visage inconnus. Cette représentation de l'environnement peut ensuite être utilisée afin de prendre des décisions proactivement.

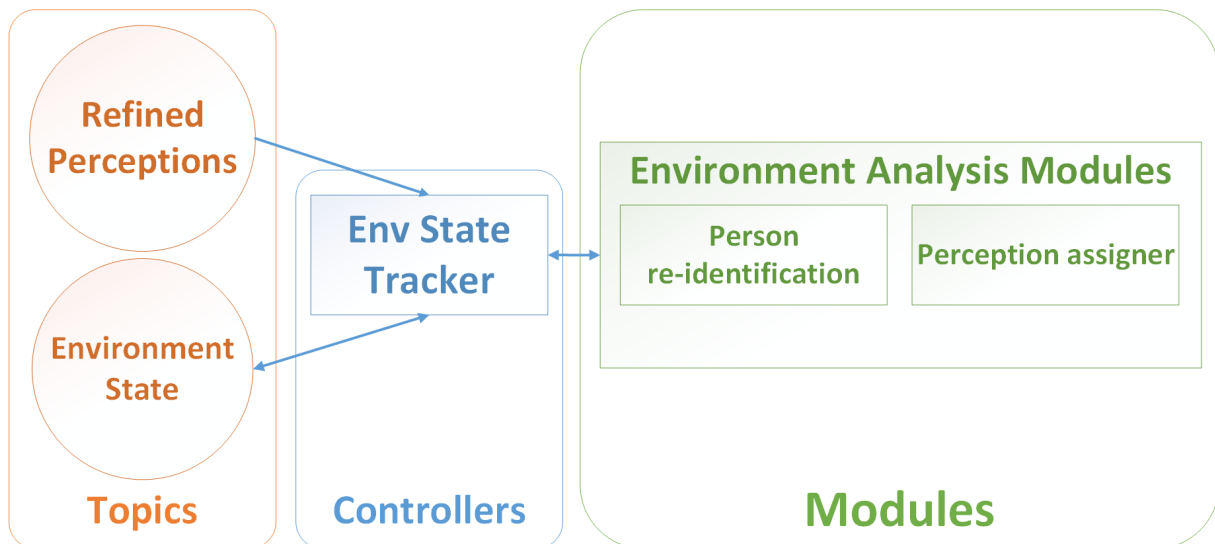


FIGURE 21 – Diagramme de la couche de représentation du système FlowAct.

6. <https://cloud.google.com/speech-to-text>

7. <https://www.deepl.com/fr/docs-api>

3.4.3.1 *Les modules d'analyse de l'environnement*

Les modules d'analyse de l'environnement servent à compléter la représentation de la scène en organisant les perceptions. Ils servent de support au module de Suivi de l'Environnement afin de délivrer la représentation la plus complète possible de la scène.

MODULE DE RÉ-IDENTIFICATION

Le suivi précis des individus dans l'environnement est essentiel pour créer une expérience utilisateur enrichie et pour permettre au robot de remplir son rôle social avec une pertinence accrue. Pour associer une identité unique à chaque utilisateur et les garder en mémoire, un module de ré-identification des personnes assiste le suiveur de l'État de l'Environnement.

La gestion visuelle des individus, grâce à leur boîte englobante et l'identifiant de suivi Deepsort, permet de suivre un individu dans le temps. Cependant, cette méthode présente des limitations en termes de reconnaissance lorsque l'individu sort et revient dans le champ de vision de l'agent. Ainsi, le module de ré-identification utilise la reconnaissance faciale lorsque l'identifiant Deepsort est inconnu, permettant de retrouver l'identité de la personne détectée si celle-ci a déjà été vue et enregistrée auparavant. Le fonctionnement complet est détaillé dans la figure 22. Afin de trier et d'assigner les boîtes englobantes, les utilisateurs sont triés et chacun possède plusieurs caractéristiques : leur dernière boîte de délimitation, leur face encodée, et les identifiants Deepsort qui leur correspondent. La nouvelle boîte de délimitation et son identifiant Deepsort sont d'abord comparés avec les identifiants Deepsort des utilisateurs à $t-1$. Si les identifiants correspondent et que le visage de la personne est extrait, la face encodée est mise à jour ainsi que la boîte englobante pour l'utilisateur correspondant. Si l'identifiant de la boîte à t ne correspond à aucun identifiant des utilisateurs à $t-1$, une reconnaissance faciale est effectuée avec les utilisateurs connus. Si la face de la boîte détectée à t correspond à un utilisateur connu, alors elle est assignée à cet utilisateur et sa face encodée est mise à jour. Si l'identifiant Deepsort et la boîte englobante ne sont toujours pas assignés, ils sont comparés avec une liste de trajectoires (boîte englobante, identifiant Deepsort, face encodée) non assignées à l'aide de l'identifiant Deepsort et de la reconnaissance faciale. Ces trajectoires non assignées permettront ensuite de créer un nouvel utilisateur lorsque celles-ci seront suffisamment mises à jour.

MODULE D'ASSIGNATION DES PERCEPTIONS

Il existe une catégorie de perceptions dépendant de la source dans l'image, qui sont les perceptions relatives à un individu dans la scène. Cela inclut la position de l'individu dans l'image, l'orientation de sa tête, l'angle de son regard, la phrase prononcée ainsi que son état émotionnel. Ces informations, extraites de l'image, arrivent sous forme de perceptions désordonnées et non nécessairement liées entre elles dans les topics de raffinement.

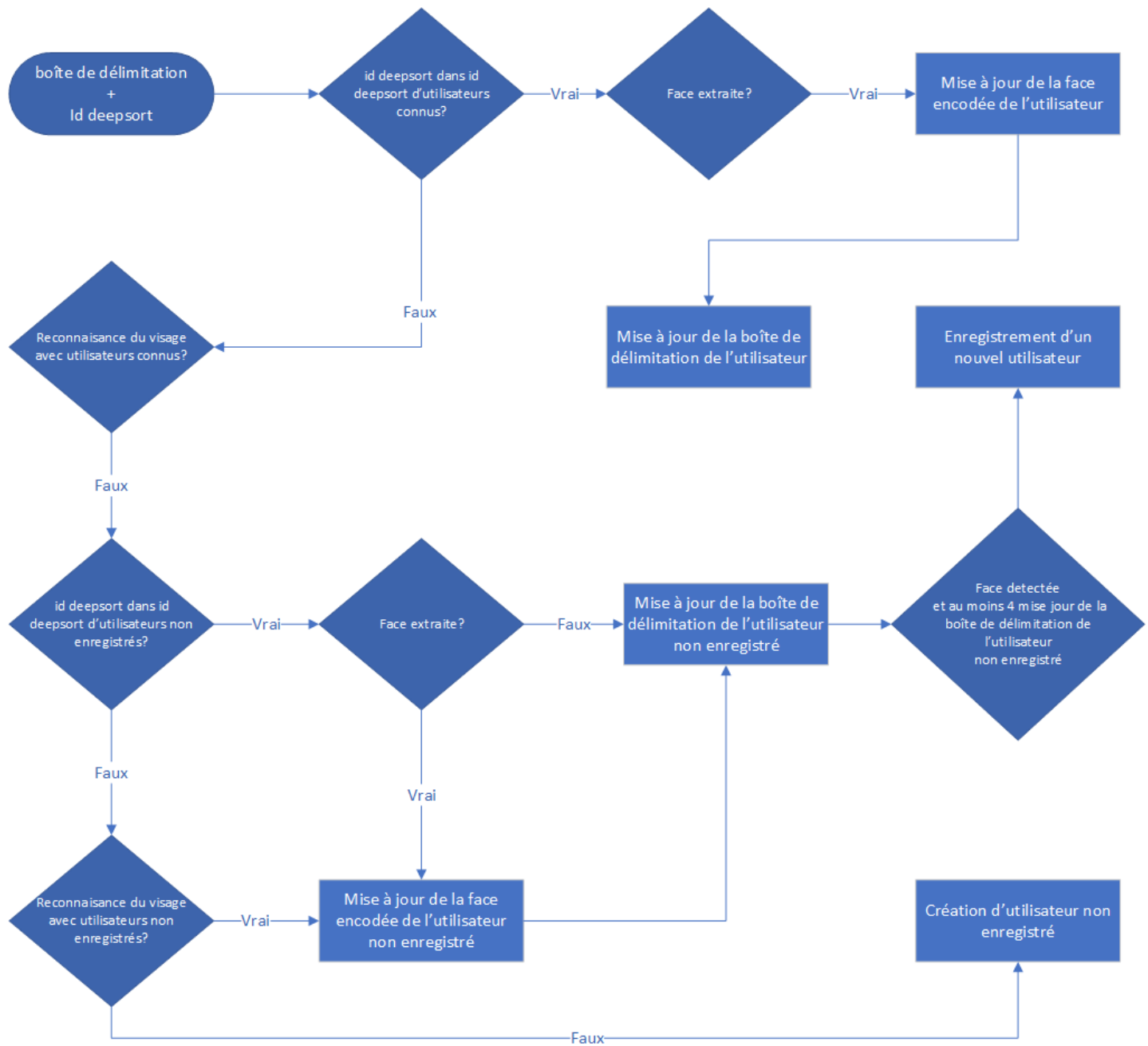


FIGURE 22 – Diagramme flowchart du fonctionnement du module de ré-identification de personne.

Dans la couche de représentation, ces perceptions sont synchronisées temporellement et toutes associées à l'individu source par des ensembles de règles. Par exemple, une phrase prononcée sera liée à l'orateur dominant détecté par le module de diarisation des locuteurs en fonction de sa temporalité. Le suivi du regard d'un individu sera attribué à une identité spécifique en fonction de la position de sa tête donnée par le module de suivi de regard et de la boîte englobante détectée par le module de détection de personne.

Certaines perceptions peuvent être préalablement associées en amont, par exemple, en passant une image avec une boîte englobante directement dans le module de suivi du regard. Cependant, pour éviter l'accumulation de latence due à l'empilement des modules et parce que certains modules ne peuvent pas s'assigner directement à l'identité d'un individu dans la scène en amont, ce module offre une flexibilité pour combiner et synchroniser les perceptions raffinées entrantes.

3.4.4 Module de décision

3.4.4.1 La planification d'action

La dynamique comportementale de l'agent est régie par le 'Planificateur d'Actions' et un ensemble de 'Modules d'Action' (figure 23). Le Planificateur d'Actions, agissant comme le centre décisionnel, prend continuellement des décisions telles que l'attente, le retour à une position d'observation, l'initiation d'une interaction ou la poursuite des activités en cours. Ces décisions sont ensuite transmises aux différents modules d'action, chacun étant conçu pour s'activer en réponse à un comportement spécifique.

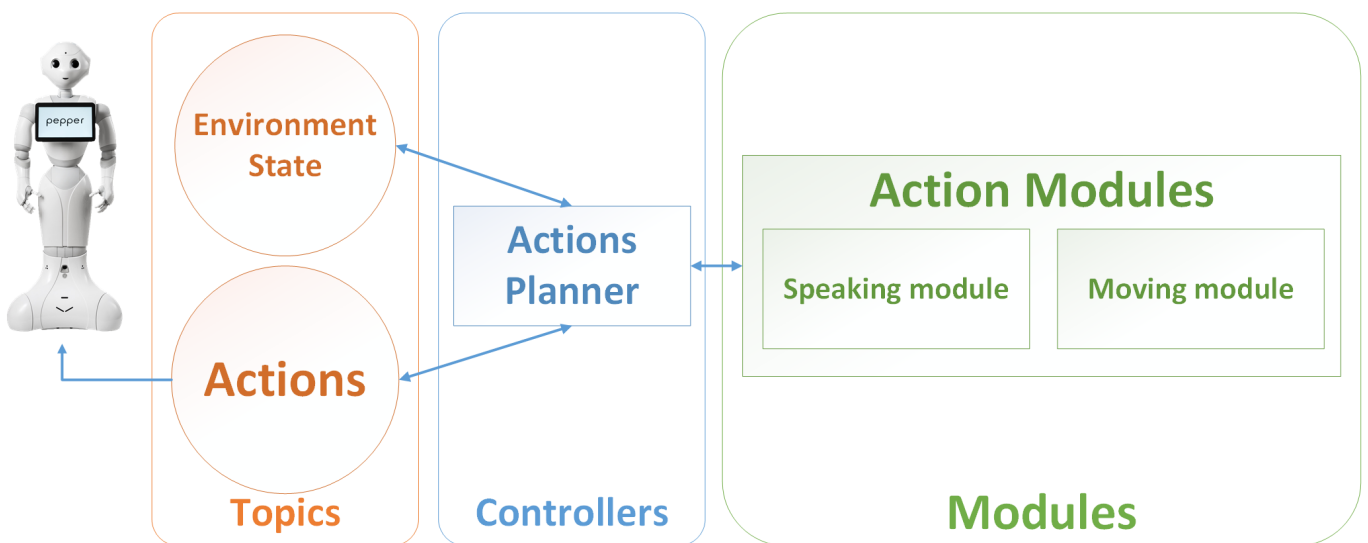


FIGURE 23 – Diagramme de la couche de décision de FlowAct.

En s'appuyant sur le cadre établi par (MOULIN-FRIER et al. 2018), nous conceptualisons les sorties du planificateur d'actions comme un 'comportement global'. Pour adhérer à ce comportement global, des 'comportements locaux' sont relayés via les modules d'action. Ces comportements locaux se caractérisent par leur nature plus granulaire et atomique. Le comportement d'engagement dans le système est régi par une série de règles basées sur le niveau du modèle de l'IAB, couplées à des demandes d'engagement implicites de l'utilisateur, telles que faire un signe avec la main.

3.4.4.2 Les modules d'actions

L'objectif principal des modules d'actions dans l'architecture Flowact est de créer une bibliothèque d'actions de haut niveau mise à la disposition du planificateur d'actions. Cela permet au planificateur d'actions de sélectionner de manière dynamique et contextuelle les actions les plus appropriées, proposées par les modules d'actions, en fonction de l'état actuel de l'environnement et des objectifs à atteindre. Les modules d'actions utilisés dans le cadre de *μdialbot* sont les prérequis pour un robot social, la parole et le déplacement. Ils fonctionnent de manière asynchrone et sont activés par le planificateur d'actions, comme suit :

MODULE DE MOUVEMENT Le module de mouvement se charge de proposer deux actions distinctes de déplacement. Premièrement, engager un individu ciblé, deuxièmement, manœuvrer le robot pour le ramener à sa position d'observation. Il est équipé d'états internes qui déterminent si le robot est stationné à son emplacement de tour de contrôle et s'il est en cours de séquence de navigation. La navigation s'effectue à l'aide de la vision et du capteur de distance. Lorsque le planificateur d'action donne l'ordre de se déplacer vers un utilisateur, l'agent s'oriente face à celui-ci, puis avance jusqu'à atteindre une distance adéquate. Pour revenir à sa position d'observation, l'agent exécute le mouvement inverse.

MODULE DE PAROLE Ce module régit le dialogue avec les utilisateurs, s'engageant dans un cycle conversationnel une fois l'engagement atteint et les paramètres de communication établis par les modules d'action.

Le module de parole a beaucoup évolué tout au long de la conception de Flowact. Initié en 2020, la première version se concentrait sur l'utilisation de Rasa⁸. C'est un framework open-source conçu pour créer des assistants virtuels et des chatbots capables de comprendre et de répondre aux requêtes des utilisateurs. Rasa se compose de deux éléments principaux. Le premier, Rasa NLU (Natural Language Understanding), assure l'analyse des messages des utilisateurs en identifiant les intentions (ce que l'utilisateur souhaite accomplir) ainsi que les entités (les informations essentielles présentes dans le message, telles que des dates, des lieux, etc.). Le second composant, Rasa Core, s'occupe de la gestion de la logique conversationnelle. Il détermine la réponse appropriée du bot en se basant sur l'intention de l'utilisateur, le contexte de la conversation, et les règles ou politiques établies. Rasa Core permet ainsi de créer des dialogues dynamiques et adaptatifs, en tenant compte de l'historique des interactions. Pour répondre à la problématique d'une conversation dans une salle d'attente d'hôpital, quatre chatbots distincts ont été développés afin de mettre en place une configuration multi-chatbots (SERBAN et al. 2017) :

- *rasa sanitary* : chargé de rappeler les règles sanitaires de l'hôpital, telles que le port du masque.
- *rasa courtesy* : utilisé pour gérer les échanges de formules de politesse.
- *rasa empathic* : responsable des réponses apportant un soutien émotionnel.
- *rasa practical* : répond à toutes les questions pratiques concernant l'hôpital, comme la localisation des toilettes.

8. [urlhttps://rasa.com/](https://rasa.com/)

Chaque chatbot a été manuellement alimenté avec des exemples simples. En plus de ces quatre chatbots orientés vers des tâches spécifiques, nous avons ajouté un cinquième chatbot à domaine ouvert : BlenderBot (ROLLER et al. 2020). Il se démarque par l'utilisation d'ensembles de données qui combinent des tâches telles que l'empathie, la personnalité et les connaissances et est disponible en trois versions selon le nombre de paramètres : 90M, 2.7B, et 9.4B. L'utilisation de BlenderBot fut motivé par le besoin de réponse dans le cas où la conversation sortirait du champ des quatre chatbots conçus. Le fonctionnement initial repose sur un modèle de règles simples. À chaque tour de dialogue, le chatbot qui propose la réponse avec le score de confiance le plus élevé pour la détection de l'intention est sélectionné. Toutefois, si ce score de confiance est inférieur à un certain seuil, indiquant que les chatbots n'ont probablement pas la réponse adéquate, c'est BlenderBot qui prend alors en charge la réponse. L'objectif était de faire fonctionner l'ensemble des chatbots lors de sessions d'expérimentation afin de recueillir un maximum de données pour les chatbots Rasa et d'utiliser les retours de satisfaction pour apprendre une politique de sélection de chatbot par apprentissage par renforcement. Mais l'arrivée des LLM a changé la donne et ces modèles se sont imposés comme une alternative plus efficace ZHAO et al. 2023 et ne demandant pas de collecte de données. Afin de guider le LLM pour répondre aux besoins conversationnels d'un agent dans une salle d'attente d'hôpital, un système de jeu de rôle est implémenté dans le prompt (NJIFENJOU et al. 2024). Un module construit les prompts selon quatre blocs :

- Instructions du Système : Ce bloc contient un ensemble d'instructions définissant précisément les spécifications de la tâche cible et le comportement global souhaité, y compris les styles de réflexion et de génération.
- Contexte Situationnel : Ce bloc regroupe les informations contextuelles qui peuvent aider le modèle à mieux accomplir la tâche. Il évolue avec le temps en fonction du flux de la conversation et peut inclure des informations sur la personnalité, des descriptions d'images et de scènes, des résumés d'échanges précédents, ou des informations provenant de sous-modules externes.
- Instructions de Réponse : Ce bloc fournit des instructions finales pour encourager le modèle de langage à répondre à l'utilisateur, en insistant sur le style d'écriture, la langue de réponse cible, et la créativité, tout en tenant compte du contexte situationnel.
- Historique de la Conversation : Ce bloc contient les messages précédents de l'utilisateur et du modèle. Les échanges peuvent être tronqués pour respecter la limite de taille des tokens du modèle, et un résumé des échanges supprimés peut être ajouté au contexte situationnel pour maintenir la cohérence.

Un module de filtrage corrige ensuite les réponses générées par le modèle de langage avant de les envoyer à l'utilisateur si elles ne sont pas valides. Le LLM utilisé est Vicuna, choisi pour son accessibilité et ses performances conversationnelles (CHIANG et al. 2023). Les réponses sont données par un appel API. La fin de la conversation dépend de la reconnaissance de mots-clés spécifiques ou du désengagement de l'utilisateur, y compris un intervalle de non-réponse de 10 secondes. Ce module met également à jour les indicateurs d'état interne, 'parle' et 'en conversation', pour refléter le discours actif du robot et son engagement dans le dialogue, pour assurer la continuité des décisions du Planificateur d'Actions.

Pour nos expériences, le LLM Vicuna, introduit en 2023, est associé au système de jeu de rôle développé par NJIFENJOU et al. (2024). Pour les tests antérieurs, une version de BlenderBot avait été employée. Le large modèle de langage a été hébergé sur les serveurs du laboratoire informatique de l'Université d'Avignon, en raison des capacités insuffisantes de mémoire vidéo de l'ordinateur portable pour charger l'intégralité du modèle. Ce modèle a été chargé sur une carte graphique A100.

3.4.5 *Résumé intermédiaire*

En résumé, cette section présente le système FlowAct, une approche préliminaire dans le domaine des systèmes perception-action continus dans un contexte robotique. L'élément central de cette étude est le développement d'une architecture ROS pour un robot d'assistance sociale, conçu pour fournir des services efficaces tout en s'engageant de manière dynamique et personnelle avec les utilisateurs. Elle se compose de trois étapes, **la perception, la représentation, la décision**. Ces étapes fonctionnent à l'aide de sous-modules organisés en contrôleurs. Cette architecture se démarque des autres propositions de système cognitif pour la HRI par sa modularité et sa gestion des flux perceptifs dans la boucle continue de perception-action. Lorsque le Planificateur d'Actions transmet un ordre aux sous-modules, ceux-ci s'exécutent de manière asynchrone, ce qui permet au planificateur de poursuivre son analyse en continu et de prendre des décisions en parallèle. Ainsi, si au cours d'une interaction l'agent perçoit un bruit suspect, tel qu'un cri, il sera capable d'interrompre l'échange pour enquêter sur la situation. Ce système d'analyse et d'action continue est indispensable pour la création d'un système capable d'initiatives afin d'agir proactivement.

3.5 EXPÉRIMENTATIONS

3.5.1 *Introduction*

En s'appuyant sur l'architecture FlowAct, conçue pour générer un comportement proactif tel que décrit précédemment, cette étude explore la perception de ce comportement par les utilisateurs, en se focalisant sur la proactivité du robot lors de scénarios d'interaction humain-robot.

L'objectif de cette étude est de valider le bon fonctionnement de l'ensemble du système et d'analyser l'expérience utilisateur dans le cadre d'une salle d'attente d'hôpital. Les expériences seront menées en deux étapes distinctes.

La première étape consiste en une expérimentation contrôlée dans un laboratoire, où des utilisateurs simulent le rôle de patients. Cette phase permet de tester et d'affiner les fonctionnalités du système dans un environnement maîtrisé, en observant les interactions et les réponses du robot social proactif.

La deuxième étape se déroulera dans un cadre réel, au sein de l'hôpital, avec de vrais patients. Cette phase vise à évaluer le système dans des conditions réelles d'utilisation, en tenant compte des variables imprévisibles et de la diversité des interactions humaines qui peuvent survenir.

Cette approche en deux temps permettra non seulement de valider techniquement le système, mais aussi d'obtenir des retours sur l'expérience utilisateur, en s'assurant que le robot puisse efficacement remplir son rôle social dans un contexte hospitalier.

3.5.2 *Expérience de laboratoire contrôlé*

3.5.2.1 *Scénario*

Afin d'évaluer l'efficacité du modèle proposé, nous avons mis en œuvre la boucle d'interaction proactive dans un cadre réel, spécifiquement conçu pour reproduire un scénario dans lequel des patients attendent leurs rendez-vous dans une salle d'attente d'hôpital. Cette mise en œuvre nous a permis de simuler un environnement authentique et de tester le modèle dans des conditions proches de la réalité clinique.

Pour émuler le rôle de patients pour cette investigation, nous avons recruté un groupe de participants diversifié, comprenant 20 individus du LIA. Parmi ces participants, on comptait 14 hommes et 6 femmes et l'âge des participants variait de 22 à 52 ans.

En ce qui concerne leur familiarité avec la robotique, seuls 3 des 20 participants avaient une expérience préalable dans ce domaine. Cette proportion limitée de participants familiarisés avec la robotique nous a permis d'évaluer le modèle dans un contexte où la majorité des utilisateurs potentiels auraient peu ou pas de connaissance préalable, ce qui est représentatif des situations réelles dans les environnements hospitaliers.

Les participants étaient positionnés à proximité de l'entité robotique, soit assis, soit debout, afin de simuler diverses postures et positions typiques des salles d'attente. Par la suite, les participants ont été instruits d'incarner l'un des comportements suivants :

- **S'engager dans une utilisation active d'appareils mobiles** : Les participants étaient encouragés à utiliser leurs smartphones ou tablettes de manière intensive, par exemple en jouant à des jeux, en naviguant sur le web ou en utilisant des applications. Ce comportement visait à simuler une situation fréquente dans les salles d'attente où les patients cherchent à passer le temps en utilisant leurs appareils électroniques ;
- **Présenter un comportement passif** : Dans ce scénario, les participants devaient maintenir une attitude d'inactivité, sans engagement apparent dans aucune activité spécifique. Ils étaient invités à adopter une posture détendue, affichant un état d'attente sans interaction avec leur environnement, ce qui est souvent observé chez les patients qui attendent leur rendez-vous sans distraction particulière ;
- **Montrer des signes d'intérêt et d'attention envers l'agent robotique** : Les participants devaient démontrer une curiosité active envers le robot. Cela incluait des comportements tels que regarder fréquemment le robot, montrer des expressions faciales d'intérêt ou se tourner légèrement dans sa direction. Ce comportement permettait d'évaluer comment le robot pouvait capter et maintenir l'intérêt des patients dans un environnement d'attente ;
- **Initier une interaction avec l'agent robotique** : Les participants étaient encouragés à chercher activement l'attention ou l'assistance du robot. Ils devaient appeler le robot en levant la main. Ce comportement visait à tester la capacité du robot à répondre de manière adéquate et utile aux sollicitations directes des patients,

simulant ainsi des interactions typiques dans une salle d'attente où les patients pourraient avoir besoin d'assistance.

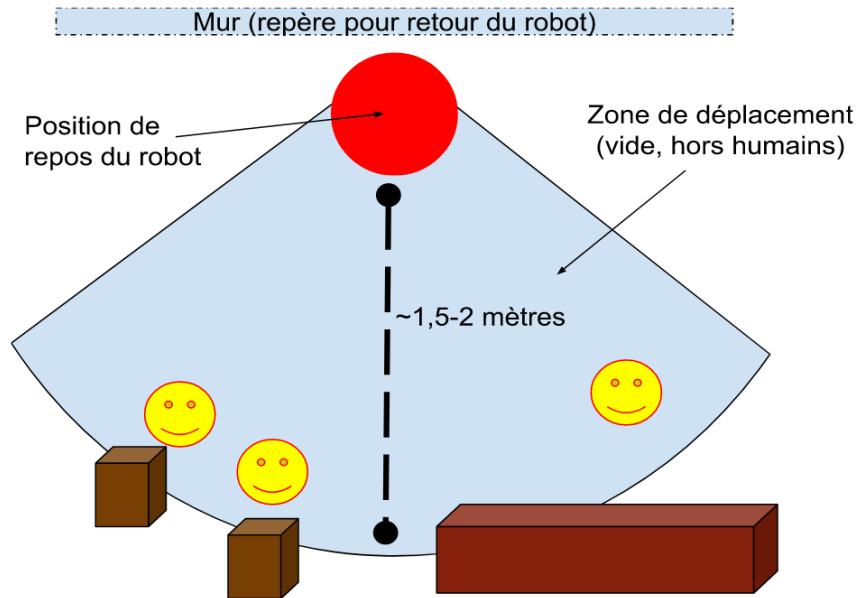


FIGURE 24 – Diagramme du schéma expérimentale de l'architecture FlowAct.

Après chaque interaction, le robot retourne systématiquement à sa position de repos (figure 24) afin de se préparer pour la prochaine interaction. Cette phase de repositionnement est essentielle pour maintenir une cohérence dans l'expérience et pour signaler aux participants la fin d'une interaction et le début potentiel d'une nouvelle. Les participants sont alors invités à reprendre l'activité qu'ils effectuaient avant d'être interrompus par l'interaction, dès que le robot se remet en position initiale. La durée de chaque interaction est déterminée par l'utilisateur. Cette flexibilité dans la durée des interactions permet de capturer une variété de réponses et de situations, enrichissant ainsi les données recueillies.

Dans chaque scénario, les participants adoptent un mélange de comportements, pouvant être passifs ou actifs vis-à-vis du robot. Par exemple, ils peuvent manifester de l'intérêt en observant attentivement le robot ou en initiant des interactions, ou au contraire, adopter une posture passive en ignorant le robot ou en se concentrant sur d'autres activités.

Un scénario est considéré comme terminé soit quand une durée prédéfinie est écoulée, soit après que chaque participant a interagi avec le robot. Cette approche permet de tenir compte des variations individuelles dans les interactions.

Chaque scénario implique la participation de deux personnes : soit deux testeurs, soit un testeur et une autre personne jouant le rôle de distraction en adoptant un comportement passif envers le robot. Cette configuration permet de simuler des environnements plus réalistes où plusieurs individus sont présents et où le robot doit gérer des interactions multiples et potentiellement concurrentes. Le rôle de la personne jouant la distraction permet d'évaluer la capacité du robot à identifier et à engager de manière optimale la personne la plus appropriée.

TABLE 13 – Questionnaire d’expérience utilisateur [UMUX](#).

Q1. Les capacités [de ce système] répondent à mes besoins.							
	1	2	3	4	5	6	7
	Pas du tout d’accord				Tout à fait d’accord		
Q2. L’utilisation [de ce système] est une expérience frustrante.							
	1	2	3	4	5	6	7
	Pas du tout d’accord				Tout à fait d’accord		
Q3. [Ce système] est facile à utiliser.							
	1	2	3	4	5	6	7
	Pas du tout d’accord				Tout à fait d’accord		
Q4. Je dois passer trop de temps à corriger les choses avec [ce système].							
	1	2	3	4	5	6	7
	Pas du tout d’accord				Tout à fait d’accord		

3.5.2.2 Évaluation

Pour valider la fonctionnalité et l’expérience utilisateur de Pepper dans le cadre d’interactions robotiques proactives, nous avons mené une évaluation sur le terrain centrée sur l’expérience utilisateur. Cette évaluation visait à valider l’efficacité de la configuration mise en place. Dans ce contexte, l’expérience est conçue comme une tâche où le robot doit démontrer un comportement d’engagement proactif envers les individus présents dans la salle. Les participants, informés des objectifs du robot, ont ensuite été invités à évaluer leur expérience en répondant à un questionnaire basé sur leurs ressentis post-expérimentation.

La tâche de l’agent lors de l’expérience est définie de la manière suivante aux utilisateurs : en tant qu’entité robotique socialement proactive, l’agent a pour mission d’initier une interaction avec les individus présents dans la pièce. Cette interaction doit se produire au moins une fois avec chaque personne considérée comme abordable selon les critères de perception du robot. Le robot doit identifier de manière autonome les opportunités d’engagement, en prenant en compte des facteurs tels que la disponibilité des individus, leur ouverture à l’interaction et la pertinence contextuelle.

Le questionnaire d’expérience utilisateur [UMUX](#), tel que décrit par l’étude de [FINSTAD 2010](#), a été utilisé pour évaluer les interactions des utilisateurs avec le système. Chaque question représente une évaluation de la composante d’utilisabilité du système. En suivant l’ordre du questionnaire [table 13](#), nous pouvons mesurer : l’efficacité, la satisfaction, la qualité globale et l’efficience. Pour un robot interactif tel que Pepper, l’efficacité pourrait être mesurée par sa capacité à engager proactivement les participants et à accomplir les tâches prévues sans erreurs ni interruptions.

L’efficacité, dans le contexte de l’expérience utilisateur, se réfère à la capacité d’un système ou d’un dispositif à permettre aux utilisateurs d’accomplir des tâches spécifiques

avec précision et exhaustivité. Elle évalue dans quelle mesure les utilisateurs peuvent atteindre leurs objectifs initiaux avec succès en utilisant le système. Pour un robot interactif tel que Pepper, l'efficacité pourrait être mesurée par son habileté à engager proactivement les participants et à accomplir les tâches prévues sans erreur ni interruption.

La satisfaction désigne le niveau de contentement des utilisateurs à l'égard de leur interaction avec un système ou un dispositif. Elle englobe des aspects subjectifs tels que le plaisir, le confort et la perception de la convivialité de l'interface. Une évaluation de la satisfaction cherche à déterminer si l'expérience globale répond ou dépasse les attentes des utilisateurs. Pour Pepper, cela pourrait inclure l'appréciation des interactions, la facilité d'utilisation et l'attitude positive ressentie après l'interaction avec le robot.

L'efficacité concerne la relation entre les ressources dépensées et les résultats obtenus lors de l'utilisation d'un système ou d'un dispositif. Elle mesure la quantité de ressources, telles que le temps, l'effort et les coûts, nécessaires pour accomplir une tâche. Un système est considéré comme efficace s'il permet aux utilisateurs de réaliser leurs objectifs avec un minimum de ressources. Dans le cas de Pepper, l'efficacité pourrait être évaluée par le temps nécessaire aux participants pour s'engager avec le robot et accomplir les interactions souhaitées.

La Métrique d'Utilisabilité pour l'Expérience Utilisateur (**UMUX**) est une échelle simple et efficace, composée de quatre questions notées de 1 à 7. Nous avons utilisé cette échelle pour évaluer comment les utilisateurs perçoivent l'utilisabilité de l'application. Bien que le questionnaire **UMUX** soit plus court que l'échelle d'Utilisabilité Système (**SUS**, BROOKE 1995) de 10 questions, il offre des résultats comparables, assurant une évaluation complète et fiable de l'expérience utilisateur.

3.5.2.3 *Considérations éthiques*

Conformément aux lois locales et aux règles de notre institution, il a été déterminé que cette recherche impliquant des participants humains ne nécessitait pas d'approbation éthique dédiée. Les participants, comprenant des universitaires et des étudiants de l'Université d'Avignon, ont donné leur consentement écrit pour faire partie de cette étude. De plus, un consentement éclairé écrit a été obtenu des individus pour la publication de toute image ou donnée identifiable qui pourrait être incluse dans des publications liées à l'expérience.

3.5.2.4 *Résultats*

Les expériences se sont déroulées sur une période de 3 semaines et environ une heure et 23 minutes ont été enregistrées au format rosbag **ROS**. La durée de l'expérience la plus longue est de 5 minutes et 6 secondes, la plus courte dure 1 minute et 27 secondes. En moyenne, les expériences durent environ 2 minutes et 29 secondes.

Les principaux résultats de nos expériences sont illustrés dans la figure 25. Les scores moyens obtenus pour chaque question étaient de 5,56, 3,25, 6,13 et 3,13 respectivement. Le score moyen **UMUX** d'environ 71 suggère que, sur une échelle de 0 à 100, l'utilisabilité globale et l'expérience utilisateur du système évalué sont bonnes. En ce qui concerne la performance de la tâche, une majorité significative des participants a estimé que l'agent remplissait efficacement son rôle. Le robot a démontré sa capacité à s'engager dans des

interactions conceptualisées dans l'environnement situé, en manifestant à la fois une réactivité aux signaux explicites des utilisateurs et une proactivité motivée par l'intérêt des acteurs pour l'agent.

L'évaluation globale de la satisfaction avec le système révèle des opinions mitigées. Bien que la majorité des participants n'aient pas éprouvé de frustration excessive, certains cas notables de mécontentement étaient principalement dus à la latence de réponse de l'agent, qui oscillait généralement autour de 5 secondes. Ce délai est attribué à la forte demande computationnelle requise par l'agent conversationnel, basé sur un grand modèle de langage. Cette latence a particulièrement affecté les participants possédant une expérience préalable en robotique, qui ont exprimé des niveaux de frustration plus élevés en raison de ces retards.

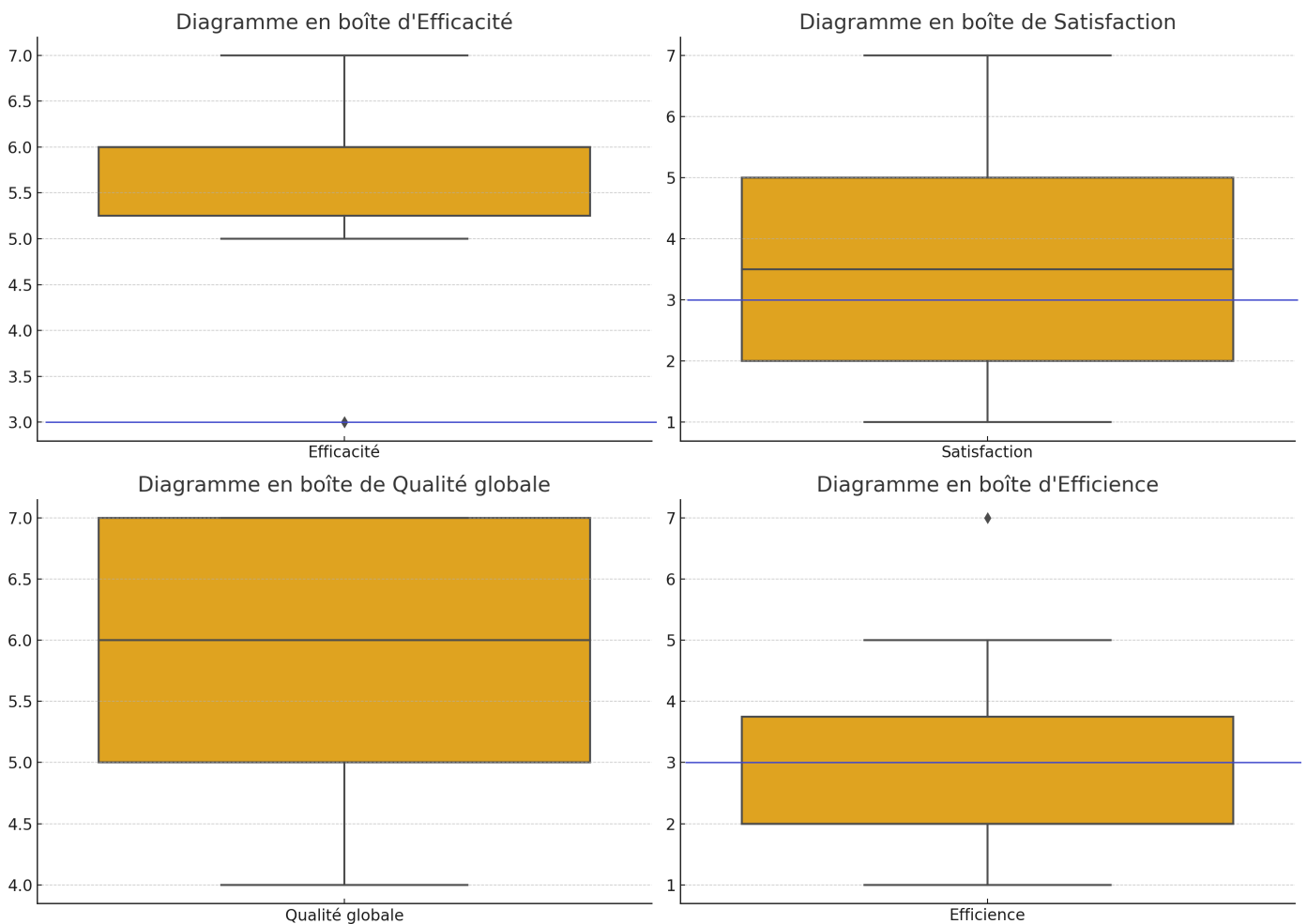


FIGURE 25 – Diagrammes en boîte des résultats du questionnaire UMUX.

En termes d'utilisabilité, le consensus était que le système était convivial, ne nécessitant aucune condition préalable spécifique pour son fonctionnement. Au cours des expériences, il a été observé que les scores d'efficience avaient tendance à augmenter dans des situations où les participants devaient maintenir un contact visuel avec l'agent plus longtemps que prévu (souvent plus de 10 secondes) ou lorsqu'ils devaient se répéter en

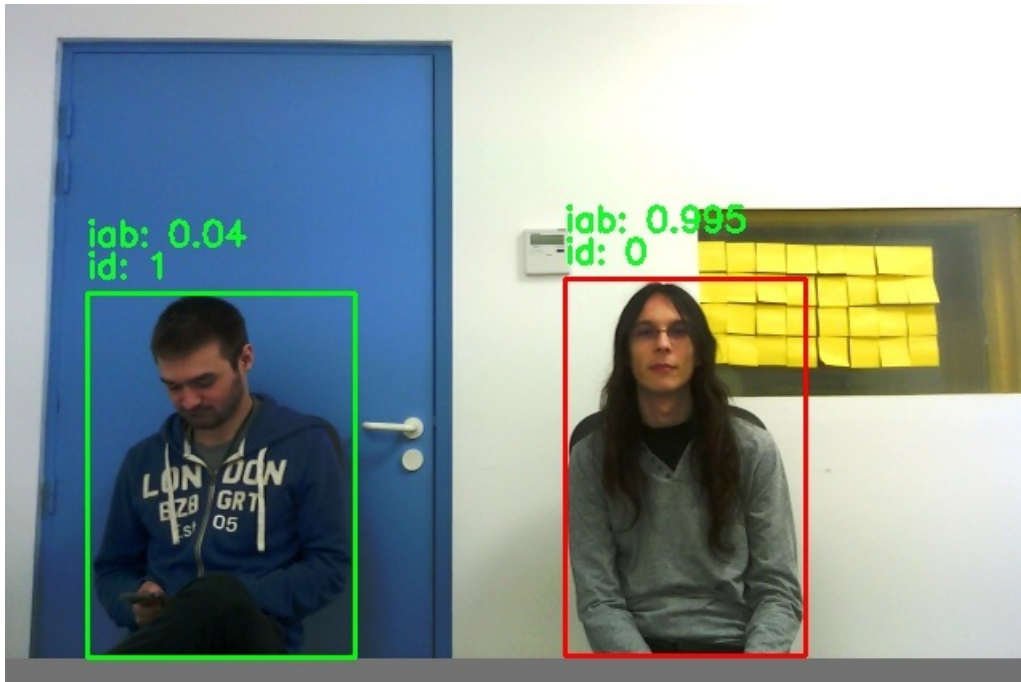


FIGURE 26 – Exemple d’une scène expérimentale, montrant l’IAB pour chaque personne détectée, avec son id. Les contours rouge indiquent une personne disponible à l’engagement.

raison de l’incapacité du robot à comprendre leur parole. Le maintien du contact visuel pour déclencher l’interaction est principalement analysé par le score de l’IAB, qui est inféré en fonction notamment du suivi du regard. Il a été noté que le module de suivi du regard perdait parfois l’angle du regard, de petits mouvements de tête permettaient de récupérer le suivi.

Quant à la compréhension de la parole, les difficultés semblent être attribuées à la moindre qualité de capture des microphones de Pepper, largement perturbée par les bruits de ses propres moteurs. Cette interférence acoustique affecte significativement la capacité du robot à interpréter correctement les commandes vocales des utilisateurs.

La figure 26 illustre un exemple de scène des expériences menées. L’utilisateur de gauche est plongé sur son téléphone illustrant une disponibilité moindre, la personne de droite est attentive au robot qui le détecte alors à l’aide de l’inférence du IAB comme potentiellement engageable.

Durant les tests réalisés, chaque essai a été mené à bien, démontrant un engagement actif et une interaction efficace avec chaque participant. Il existe deux cas spécifiques, dans lesquelles le robot a pris l’initiative d’engager une interaction avec les individus avant même qu’ils ne manifestent un signe réactif. Les utilisateurs ont perçu cette initiative proactive du robot comme un signe d’intelligence et d’anticipation, car elle est survenue en réponse à leur intérêt pour le robot qui se situait dans le dialogue, perçu avant même le signal conventionnel d’engagement.

Cette capacité du robot à anticiper et à initier le contact a été globalement bien accueillie, ajoutant une dimension de fluidité et de réactivité à l’interaction. Cependant, il a été observé que le modèle de langage large utilisé par le robot a généré quelques hallu-

cinations, c'est-à-dire des réponses inappropriées ou sans rapport direct avec le contexte. Ces occurrences n'ont pas systématiquement été interprétées de manière négative par les utilisateurs ; au contraire, elles ont parfois été perçues avec amusement, ajoutant une touche d'imprévisibilité qui a pu rendre l'interaction plus intéressante et divertissante.

3.5.3 *Expérience à l'hôpital Broca d'AP-HP*

3.5.3.1 *Introduction*

Dans le cadre du projet ANR μ Dialbot, une série d'expériences a été réalisée sur une période de trois jours à l'hôpital Broca d'AP-HP. L'objectif de ces expériences était de tester l'intégralité du système d'interaction multimodal proactif dans un environnement clinique. Par rapport à la configuration en laboratoire, des ajustements ont été apportés pour l'adapter au contexte hospitalier. Les utilisateurs étaient des patients de l'hôpital Broca, accompagnés par un membre du pôle de gériatrie. Afin de garder l'anonymat des patients et respecter la politique de RGPD du projet ANR, aucune donnée n'a été envoyée à des systèmes non propriétaires, l'utilisation de la reconnaissance automatique de la parole de Google a été remplacée par un modèle Whisper (RADFORD et al. 2022).

Le scénario de l'expérience est similaire à l'expérience en laboratoire, seuls différents le cadre physique et la nature des participants à l'expérience, ainsi que le fait qu'ils sont moins guidés pour la réalisation de l'interaction.

3.5.3.2 *Évaluation*

Après discussion avec nos collègues psychologues de l'hôpital Broca, l'analyse des expériences est faite à travers deux questionnaires complémentaires, le questionnaire AES (TARIMAN et al. 2009) et le questionnaire SUS (BROOKE 1995).

N°	Items
1	À quel point avez-vous trouvé ce robot facile d'utilisation (parler, se comporter, agir) ?
2	À quel point les demandes et les suggestions du robot étaient-elles compréhensibles ?
3	À quel point avez-vous apprécié l'utilisation (parler, se comporter, agir) de ce robot ?
4	À quel point ce robot vous a-t-il été utile pour (a. vous accueillir / b. favoriser des interactions sociales sans risques sanitaires / c. vous aidez à préparer les consultations / d. vous orienter et vous guider / e. vous divertir) ?
5	Le temps consacré par ce robot à (a. vous accueillir / b. favoriser des interactions sociales sans risques sanitaires / c. vous aidez à préparer les consultations / d. vous orienter et vous guider / e. vous divertir) était-il acceptable ?
6	Comment évaluez-vous votre satisfaction générale de ce robot ?

FIGURE 27 – Questionnaire d'expérience utilisateur AES.

Le questionnaire AES (Figure 27) est conçu pour mesurer l'acceptabilité perçue des nouvelles technologies par les utilisateurs. Il se concentre sur deux dimensions principales : la facilité d'utilisation perçue et la satisfaction générale. Chaque item du questionnaire est noté sur une échelle de 1 à 5, où 1 représente le score le plus négatif et 5 le score le plus positif. Ce questionnaire permet de recueillir des données quantitatives sur l'acceptabilité du robot, offrant ainsi des indications sur les aspects de conception à améliorer.

Le questionnaire SUS (Figure 28) permet lui d'évaluer la perception des participants concernant les performances du robot dans le contexte de l'interaction. Le questionnaire SUS, à l'image du questionnaire UMUX utilisé précédemment, est un outil de mesure standardisé utilisé pour déterminer l'utilisabilité d'un système. Il se compose de dix affirmations évaluées sur une échelle de Likert en cinq points, allant de "Tout à fait en désaccord" à "Tout à fait d'accord". Cette échelle permet de recueillir des données quantitatives sur différents aspects de l'interaction, tels que la facilité d'utilisation, la complexité perçue, la confiance de l'utilisateur et la courbe d'apprentissage. Les scores obtenus sont ensuite convertis en un score global sur 100, facilitant ainsi la comparaison et l'interprétation des résultats.

N°	Items
1	J'aimerais utiliser (parler, se comporter, agir) ce robot à l'hôpital le plus fréquemment possible.
2	Je trouve les conversations avec ce robot inutilement complexes.
3	Je pense que ce robot est facile à utiliser (parler, se comporter, agir).
4	Je pense que j'aurais besoin d'aide pour être capable d'interagir (échanger, parler) avec ce robot.
5	J'ai trouvé que les différents services (fonctions) de ce robot ont été bien pensés (conçus).
6	Je pense qu'il y a trop d'incohérences chez ce robot.
7	J'imagine que la plupart des gens seraient capables d'apprendre à se servir de ce robot très rapidement.
8	J'ai trouvé très difficile de parler et de me comporter naturellement avec ce robot.
9	Je me suis senti très en confiance en me servant du robot.
10	Je pense qu'il y a beaucoup de choses à apprendre pour pouvoir bien se servir de ce robot (parler, se comporter, agir).

FIGURE 28 – Questionnaire d'expérience utilisateur SUS.

3.5.3.3 Considérations éthiques

La tenue des expériences à l'hôpital Broca avec des patients a été validée par le comité éthique de chaque acteur du projet μ DialBot. L'ensemble des données des patients est anonymisé après la collecte et aucun module de l'architecture ne transmet de données sur des serveurs extérieurs au projet. Une fois anonymisées, les données sont transférées

de façon sécurisée sur un serveur de l'INRIA, où l'ensemble des partenaires peut ensuite les récupérer avec des procédures d'accès sécurisées.

3.5.3.4 Résultats

Les expériences ont été menées à l'hôpital Broca sur une période de trois jours consécutifs. Au total, onze patients ont participé à l'expérience, dont quatre hommes et sept femmes, pour un total de 13 interactions. L'âge moyen des participants était de 75 ans, avec le participant le plus jeune ayant 52 ans et le plus âgé 89 ans. Parmi ces onze personnes, sept avaient déjà eu des interactions antérieures avec des robots. La durée moyenne des interactions est de 4 minutes et 2 secondes, le total est de 58 minutes d'enregistrement.

Malgré nos efforts et pour des raisons difficilement maîtrisables liées au contexte hospitalier, ce panel est bien sûr trop petit pour tirer des conclusions pertinentes et solides des observations collectées. Toutefois, dans le cadre de cette thèse, il nous permet de poser le cadre opérationnel de l'étude et sa mise en œuvre pratique. Et d'autres séquences d'expérimentations devront venir compléter ce premier jeu très préliminaire. En attendant, nous procédons à l'analyse des résultats actuels, même insuffisants pour en tirer quelques enseignements sur l'état actuel du système et la possibilité de quelques améliorations immédiates avant de nouvelles interactions en situation réelle.

Les résultats disponibles du questionnaire utilisateur [SUS](#) sont présentés dans la figure [29](#). Le score [SUS](#) moyen obtenu lors de l'évaluation est de 59, indiquant un niveau d'utilisabilité modéré. Selon le cadre de notation du SUS, les scores proches de 51 sont considérés comme « passables » ou « moyens », tandis que les scores de 71 sont généralement interprétés comme « bons » [BANGOR, KORTUM et MILLER 2009](#). Un score de 59 suggère donc que le système évalué présente des défis notables en matière d'utilisabilité. Ce score implique que les utilisateurs pourraient rencontrer des difficultés lors de l'interaction avec le système, ce qui pourrait potentiellement affecter leur satisfaction et leur engagement global. Les réponses à la première question sur l'acceptation générale montrent que les utilisateurs ne sont pas encore prêts à utiliser le robot de manière régulière. Cela, malgré le fait que la complexité perçue soit relativement faible (question deux) et que la facilité d'utilisation soit reconnue (question trois). Les patients se sentent capables d'utiliser le robot de manière autonome, mais estiment que l'intégration des services pourrait être améliorée. Ce sentiment est compréhensible, compte tenu de la lenteur observées des déplacements du robot et de la latence de réponse du module de langage. Bien que les diverses erreurs inhérentes aux [LLM](#) aient été présentes, les participants n'ont pas trop pénalisé les incohérences dans certains dialogues (question six). Globalement, les participants estiment que ce système d'interaction humain-robot est simple et efficace en terme d'utilisabilité, capable d'instaurer la confiance chez le patient, mais ils ne sont pas enclins à l'utiliser ou à y être confrontés fréquemment.

Le récapitulatif des réponses des patients au questionnaire [AES](#), figure [30](#), confirme une perception globalement positive de la facilité d'utilisation du système. Les réponses à la question 2 indiquent une clarté satisfaisante des demandes et des suggestions émises par le robot. Certains patients apprécient son utilisation, tandis que d'autres se montrent plus réservés. Certains patients trouvent le robot utile, mais ils ne sont pas entièrement

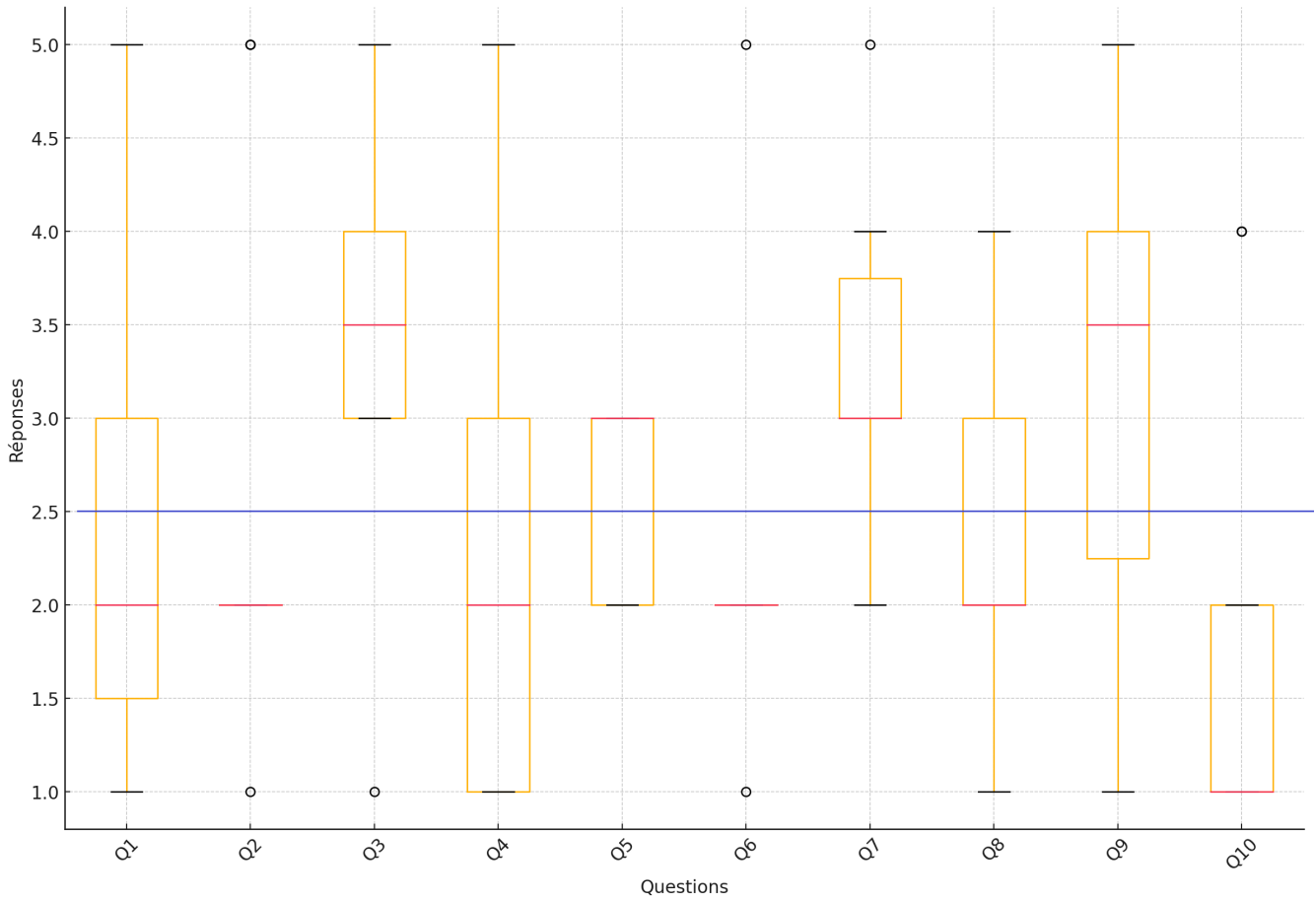


FIGURE 29 – Diagrammes en boîte des résultats du questionnaire [SUS](#).

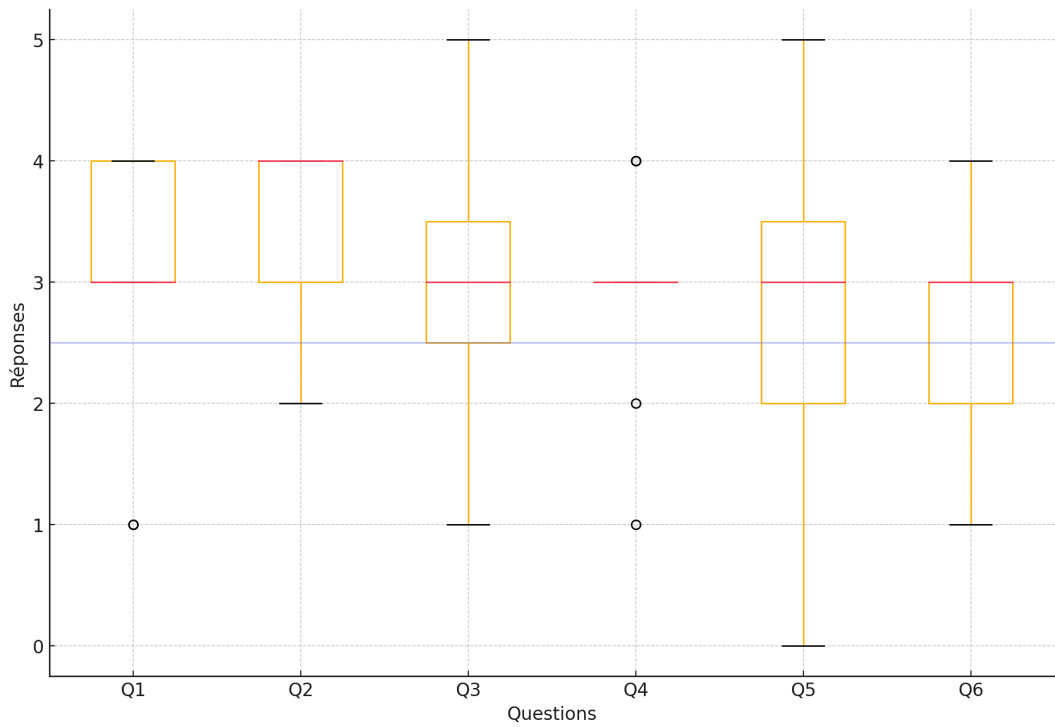


FIGURE 30 – Diagrammes en boîte des résultats du questionnaire [AES](#).

satisfaits de son état actuel. Par ailleurs, les retours concernant la question du temps consacré au robot sont relativement neutres, montrant une diversité d'opinions parmi les patients. Le robot ne semble également pas avoir convaincu les patients de son utilisation dans le cadre d'un hôpital, à l'instar du questionnaire précédent.

3.5.3.5 *Analyse des résultats*

Lors des expériences, la détection d'une valeur élevée de l'IAB était privilégiée pour assurer l'engagement des patients. Cependant, en cas de difficulté à s'engager, les patients avaient la possibilité de lever la main pour signaler leur problème. Ainsi, seules 23% des interactions ont été initiées par la détection d'une valeur élevée de l'IAB, tandis que les 77% restantes ont été déclenchées à la demande des patients, par le biais d'un geste de la main.

Contrairement aux expériences menées en laboratoire, celles-ci se sont déroulées dans trois endroits différents de l'hôpital, chacun ayant des conditions lumineuses variables. Ces conditions de luminosité ont rendu la prédiction de l'IAB plus complexe, car le module de prédiction de l'orientation du regard, qui alimente le module principal, est particulièrement sensible à la lumière. Ainsi, après analyse des données, nous observons qu'une bonne détection de l'IAB pour un engagement réussi n'a correctement fonctionné que dans moins de la moitié des interactions. De plus, le module de détection des personnes a également été parfois mis à rude épreuve, la captation du visage étant également très sensible aux conditions de luminosité. Ces observations mettent en évidence le manque de robustesse du modèle d'IAB utilisé.



FIGURE 31 – Exemple d’une séquence d’engagements successif.

La figure 31 illustre une séquence de l’engagement successif des patients. On observe, en suivant l’ordre de présentation des images de gauche à droite et de haut en bas, que le patient situé à droite n’est initialement pas détecté correctement, ce qui empêche toute interaction dans un premier temps. Le robot commence dans un mode d’observation, où il détecte un signal de l’utilisateur, en l’occurrence un appel de la main, pour initier l’engagement. Une fois la conversation démarrée, le robot prendra la décision de la continuer jusqu’à ce que le module de dialogue considère que la conversation est terminée. Ensuite, le robot retournera au mode d’observation, attendant qu’un nouveau signal soit émis par la personne située à droite, afin de l’engager à son tour et d’entamer une nouvelle conversation. Ce schéma ‘poste d’observation, engagement, interaction, retour au poste d’observation’ représente la boucle standard de l’objectif proactif de l’agent dans une salle d’attente d’hôpital. Il est intéressant de constater que la personne en blouse blanche à l’arrière-plan est bien détectée, mais n’est pas considérée comme engageable dans les photos 2 et 3, probablement en raison de son comportement sur son téléphone,

qui ne la désigne pas comme une cible potentielle. Sur la dernière photo, on observe qu'elle se rapproche de la personne engagée et est alors détectée comme engageable, par sa valeur d'IAB, ce qui semble cohérent car elle regarde directement l'agent en venant prêter assistance à un utilisateur.

Patient : Bonjour, moi c'est X et vous ?
Pepper : Bonjour, je suis Pepper, le robot conversationnel de l'hôpital Broca. Comment puis-je vous aider aujourd'hui ?
Patient : Je suis trop en avance pour mon rendez-vous, que dois-je faire en attendant ?
Pepper : Vous êtes en avance pour votre rendez-vous. Vous pouvez prendre un moment pour vous relaxer et attendre que le personnel médical vous conduise aux salles de consultation à l'heure prévue. Cela peut prendre entre 5 minutes et une demi-heure, en fonction du nombre de consultations pour la journée.
Patient : Mais comment je peux faire pour me relaxer ? Qu'est-ce que je peux faire pour me relaxer en attendant ?
Pepper : Vous pouvez prendre un moment pour vous relaxer en attendant votre consultation en feuilletant un livre ou en écoutant de la musique. Le personnel médical vous conduira aux salles de consultation à l'heure prévue.
Patient : Est-ce que vous pouvez mettre un peu de musique ?
Pepper : Je peux vous proposer de mettre de la musique, mais étant un robot, je n'ai pas accès à un système de musique. Pouvez-vous me dire si vous aimeriez que je parle de musique avec vous en attendant la consultation ?
Patient : Je veux bien.
Pepper : Comment allez-vous aujourd'hui ?

FIGURE 32 – Exemple de conversation entre un patient et le robot Pepper.

Le dialogue 32 est un exemple de conversation entre un patient et l'agent. Pepper montre une bonne compréhension des requêtes de l'utilisateur et offre des suggestions pour passer le temps en attendant le rendez-vous médical. Lorsque l'utilisateur demande à Pepper de mettre de la musique, le robot explique ses limitations techniques (absence d'accès à un système de musique) mais propose une solution alternative en offrant de discuter de musique. Cette réponse montre que le modèle de langage est capable de gérer les limitations technologiques tout en maintenant une interaction engageante. Cependant, il répondra à côté lorsque le patient lui confirmera son envie de discuter de musique. Cette difficulté à assurer un dialogue cohérent au-delà de quelques tours de parole semble avéré et doit nous amener à améliorer la prise en compte de l'historique dans le chatbot, qui malgré la connaissance de l'ensemble des échanges précédents ne parvient malgré tout pas à poursuivre une discussion jusqu'à son terme logique.

Lors de ces expériences, les conversations entre l'agent et le robot ont été difficiles en raison de la faible intensité et de la clarté insuffisante des paroles des patients. Parfois, une simple répétition suffisait, mais souvent la phrase reconnue par l'ASR différait grandement de celle réellement prononcée par l'utilisateur. Cela a compliqué les interactions avec les patients, soulignant la nécessité d'améliorer le système de captation sonore de Pepper pour une utilisation efficace avec les personnes âgées.

3.6 CONCLUSION

Les contributions présentées dans ce chapitre couvrent la conception d'un système d'interaction humain-machine dans le cadre d'une salle d'attente. Le flot de perception débute par l'extraction et le traitement des perceptions brutes. Cette étape initiale est cruciale pour la compréhension de l'environnement par le robot et la base de toute interaction subséquente. Ensuite, le raffinement des perceptions approfondit l'analyse des différentes modalités, telles que la parole, la vision et la diarisation multimodale du locuteur. Des techniques avancées sont proposées pour analyser l'engagement et l'acceptation de l'engagement dans un contexte HRI, démontrant ainsi l'importance de ces perceptions raffinées pour des interactions plus naturelles et fluides entre les humains et les robots.

La présentation du modèle FlowAct permet de lier ce flux de perception aux actions de l'agent en introduisant une architecture modulable pour l'interaction humain-robot. Ce modèle englobe plusieurs aspects clés, incluant la perception, la modélisation de l'environnement, la prise de décision et la planification d'action. Le fonctionnement général et les composants du modèle ont été détaillés pour illustrer comment FlowAct peut améliorer l'efficacité et la cohérence des interactions. Notamment en permettant à la boucle d'interaction principales d'être ouverte en permanence. En particulier le démarrage d'une interaction dialogique n'interrompt pas la perception active, et permet au module d'action du robot d'envisager la poursuite du dialogue en même temps que la possibilité de passer à une tâche plus urgente (par exemple suite à la détection d'un cri ou de l'appel d'un soignant etc). Cette dernière possibilité n'a pas été testée dans nos premiers scénarios pour les expériences réels, mais devrait l'être dans les prochaines configurations.

Les expérimentations présentées ici ont permis de valider les approches et les modèles développés dans des contextes variés, allant d'expériences contrôlées en laboratoire à des scénarios réels comme l'hôpital Broca. Ces expériences ont fourni des résultats concrets et des retours d'expérience, confirmant en première instance l'applicabilité et les avantages des solutions proposées dans le cadre d'une salle d'attente d'hôpital.

En conclusion, ce chapitre a mis en avant des contributions dans le domaine de l'HRI, en particulier par le développement et l'intégration de modules robotiques et d'un modèle cognitif plus global de gestion des perceptions afin de prendre des décisions proactives. Les résultats obtenus démontrent le potentiel des approches proposées pour améliorer les interactions entre les humains et les robots, ouvrant ainsi la voie à des applications plus complexes où le robot interroge constamment son environnement pour réagir proactivement au moindre signe social d'un utilisateur de la scène.

Quatrième partie

Conclusion, discussion et perspective

CONCLUSION, DISCUSSION ET PERSPECTIVES

4.1 CONCLUSION

Aujourd'hui, la capacité des robots à interagir de manière proactive avec leur environnement est une étape cruciale dans la conception des robots sociaux. La proactivité se manifeste de différentes manières, notamment par l'anticipation des besoins des utilisateurs. Par exemple, un robot domestique pourrait commencer à nettoyer une maison désordonnée sans qu'on le lui demande, et un robot assistant dans un hôpital pourrait apporter de l'eau à un patient avant qu'il ne le réclame lorsque la température est élevée.

La proactivité se traduit aussi par la prise d'initiatives : dans un environnement de travail, un robot pourrait offrir son aide à un collègue en difficulté, et dans une maison de retraite, organiser des activités adaptées aux préférences des résidents. Elle inclut également l'adaptation au contexte : un robot social pourrait ajuster son comportement en fonction de l'humeur de la personne, en adoptant un ton plus doux et en offrant des mots de réconfort si son interlocuteur semble stressé. De plus, elle implique une interaction naturelle et continue : un robot pourrait participer activement à la vie quotidienne d'une famille, en engageant des conversations pertinentes, en rappelant les événements importants comme les anniversaires et en proposant des idées de cadeaux. Une telle approche englobe aussi l'apprentissage et l'amélioration continus.

Ainsi, cela nécessite une multitude de compétences pour fonctionner efficacement. Pour un robot social, cette proactivité repose sur la représentation de la scène capturée par les capteurs de l'agent. Cet environnement doit être aussi complet que possible afin de permettre au robot de prendre des décisions proactives.

Cette thèse s'est attachée à répondre à la conception d'un robot multimodal proactif, partant de l'extraction des perceptions à la phase de prise d'action. Elle s'est dotée d'un cadre applicatif correspondant à l'interaction proactive dans une salle d'attente d'un hôpital.

4.2 CONTRIBUTIONS

Au regard du premier objectif de nos travaux correspondant au **O1. Développement d'un système d'interaction multimodal proactif dans le cadre d'une interaction humain-robot** posé dans le premier chapitre, il est possible de résumer les contributions de la manière suivante :

- Un modèle de fusion temporelle audio-visuelle pour la diarisation de locuteurs multi-utilisateurs, basé sur une modélisation de la fusion des indices audio et visuels par coïncidence spatiale de localisation des sources visuelles et sonores. Le modèle est relativement léger au niveau calculatoire et présente une capacité d'adaptation à différentes conditions acoustiques sans phase d'entraînement. L'ensemble est testé directement sur un jeu de données enregistré sur le robot Pepper

et aussi sur la base de données CAV3D. La **DER** sur les deux jeux de données, respectivement 19,27 et 12,06 indique des performances acceptables dans un contexte de **HRI**.

- La définition de concept de l'**IAB** pour la **HRI**, la croyance d'acceptation de l'interaction qui répond à la question "Quelles sont les chances que mon interaction soit acceptée par l'utilisateur ciblé". L'analyse de ses caractéristiques dominantes et leur modélisation révèlent que l'orientation du regard est déterminante pour la prédiction de l'**IAB**, la position du corps joue également un rôle dans une moindre mesure.
- Une architecture d'interaction humain-robot multimodale proactive, FlowAct, avec un flux continu de perception et des sous-systèmes d'action modulaires. Elle se compose de trois étapes, la perception, la représentation, la décision. Ces étapes fonctionnent à l'aide de sous-modules organisés par des contrôleurs ;
- L'implémentation du système d'interaction complet en **ROS** et sur le robot social Pepper. Les détails sur la conception et l'implémentation de chaque module du dispositif.

Les contributions concernant le deuxième objectif qui consiste à **O2. Tester et évaluer cette architecture dans un scénario de salle d'attente d'hôpital** sont présentées ci-dessous :

- Une expérience d'exploration contrôlée, d'interaction humain-robot proactive qui simule les comportements dans une salle d'attente d'hôpital. Cette expérience inclut une évaluation de l'expérience utilisateur via le questionnaire UMUX et une analyse du fonctionnement en temps réel de l'implémentation de FlowAct. Le score du questionnaire UMUX, 71 indique une expérience utilisateur satisfaisante ;
- La conduite d'un test du dispositif dans un environnement réel, à l'hôpital Broca, ainsi qu'un retour sur l'utilisabilité et l'acceptation du dispositif par les patients de l'hôpital à travers les questionnaires SUS et AES. Le score d'expérience utilisateur est de 59, indiquant un niveau modéré d'utilisabilité.

4.3 DISCUSSIONS

L'un des sujets majeurs de cette thèse est l'implémentation d'un modèle proactif d'interaction humain-robot dans un cadre hospitalier. La proactivité est un concept précis, elle est décrite de manière précise par GROSINGER (2022). Cette thèse s'appuie sur sa définition de la proactivité : la capacité d'initier de manière autonome des actions anticipatrices, fondées sur le raisonnement, visant à influencer les individus et/ou leur environnement. Plus précisément, il est possible de placer cette étude dans la catégorie de la proactivité avec un objectif donné. En ce sens, la disposition du robot à interagir avec les patients, guidée par l'analyse de leur comportement, permet de simuler un comportement proactif orienté vers l'objectif d'interagir avec le plus grand nombre de personnes réceptives à l'interaction.

Il existe un cadre théorique général de la proactivité, décrit dans la sous-section 2.5.3, portant sur la modélisation de la proactivité. Toutefois, il a été décidé de ne pas utiliser directement ce cadre dans cette étude. Bien que le framework de l'*Equilibrium Maintenance* présente des intérêts pour la représentation de la proactivité, il comporte également plu-

sieurs défauts tels que son utilisation de la logique floue et la complexité du modèle probabiliste qui en découle. Dans notre cas, le système proactif visé a pour objectif de faciliter l'interaction proactive dans une salle d'attente. Étant donné que les cas d'utilisation sont restreints aux personnes disponibles et/ou recherchant activement une interaction, un simple système de règles du premier ordre est suffisant pour modéliser le niveau décisionnel de cette proactivité dans un premier temps. L'accent est mis sur l'agencement des hypothèses avant la prise de décision, afin de permettre un comportement proactif très cohérent pour l'instant, faute d'être encore subtil.

Ainsi, cette thèse propose une méthodologie pour construire un système d'interaction multimodal proactif. Ce système traite les perceptions brutes du robot et les transforme en ordres d'action, organisant l'ensemble en modules interconnectés dirigés par des contrôleurs jouant le rôle de chefs d'orchestre. Les flux de perceptions sont triés par types, distinguant les perceptions brutes des perceptions raffinées. Grâce à cette architecture modulaire et évolutive, le système est adaptable et extensible. Les seules limites étant l'accroissement de la latence due à l'ajout de nouvelles fonctionnalités, ainsi que leur niveau de performance, bien sûr, et son impact sur la prise de décision. Enfin, bien que cette implémentation soit réalisée sur le robot Pepper, elle est directement transférable à tout agent robotique, à condition que l'extraction des perceptions et l'envoi des actions puissent être effectués via [ROS](#).

4.4 LIMITATIONS

Dans le cadre d'un projet ambitieux financé par l'Agence Nationale de la Recherche, cette thèse aborde des enjeux cruciaux en matière d'interaction humain-robot. Notamment, elle examine l'utilisation du robot humanoïde Pepper, qui, malgré ses qualités anthropomorphiques favorisant l'acceptation et sa multitude de capteurs, présente des limitations significatives pour la recherche. Le principal défaut de Pepper est l'absence de lidar, un dispositif de télédétection utilisant des faisceaux laser pour mesurer les distances et créer des modèles tridimensionnels détaillés des surfaces et des objets. Cet instrument est souvent complété avec un algorithme de *Simultaneous Localization and Mapping* (SLAM) pour permettre au robot de se déplacer efficacement dans son environnement. En l'absence de lidar, les déplacements de Pepper sont très restreints, ce qui nous a conduits à simplifier sa navigation en utilisant la vision pour détecter les êtres humains et effectuer de simples marches arrière pour son repositionnement. Un passage du robot Pepper au robot Ari de Pal Robotics était initialement prévu au cours du projet. Cependant, en raison des différences dans les modalités de déplacement et des problèmes de résonance du moteur, qui perturbaient la qualité de la capture audio en se propageant jusqu'au microphone, il a été décidé de persévérer avec Pepper.

L'un des objectifs de cette thèse était aussi d'améliorer la capacité du robot à prendre des décisions pour gérer les tours de parole de manière plus naturelle, alternant entre des interactions en groupe et des dialogues en face à face selon les besoins spécifiques. Les interactions en groupe représentent un domaine particulièrement complexe dans l'interaction humain-robot, nécessitant une analyse minutieuse des dynamiques de prise de parole et une détection précise de chaque interlocuteur afin que le robot puisse répondre de manière opportune et offrir des réponses de qualité. Cependant, ce type d'interaction

présente des défis techniques importants, notamment en ce qui concerne la latence de réponse de l'agent et les complexités liées au déplacement et au positionnement du robot Pepper dans un espace social pour les interactions *multiparty*. Pour cela, de nombreux problèmes pratiques actuels devront être réglés, par exemple, la gestion de l'écho de sa propre voix/synthèse. Lorsque le robot parle, ses microphones captent le signal sonore émis directement depuis ses haut-parleurs, ce qui entraîne un traitement dans le traitement audio. Ainsi, nous avons réalisé que les modules de détection d'écho sont encore complexes à mettre à œuvre et entraînent une latence supplémentaire dans un dispositif déjà déficient sur cette question. Aussi, dans notre cas, pour éviter que les réponses de Pepper ne soient envoyées par erreur au module de reconnaissance automatique de la parole et éventuellement au module de parole, il a été purement et simplement décidé de désactiver l'écoute des microphones de Pepper pendant qu'il parle. Cette solution applicable dans un cadre d'interaction directe (face-to-face) doit être revue pour des interactions en groupe (*multiparty*). De même qu'une séparation de source est alors nécessaire pour attribuer les segments de parole au bon interlocuteur. Autant d'enjeux qui présentent des solutions actuellement, mais qu'il n'est pas immédiat d'intégrer à une plateforme comme la nôtre sans remettre en cause son fonctionnement général (ne serait-ce que par leurs besoins en mémoire graphique).

Un autre problème majeur de Pepper réside dans la qualité et le champ de vision des caméras du robot. Ces caméras, extrêmement sensibles aux conditions d'éclairage, ne parviennent pas à capturer l'ensemble de la scène. Une solution convenable pour remédier à ce défaut aurait été l'installation d'une caméra supplémentaire greffée sur la tête du robot. Cependant, le temps nécessaire pour mettre en place ce dispositif n'a pas permis de concrétiser cette amélioration.

Face à ces défis techniques, nous avons fait le choix stratégique de nous concentrer sur l'interaction en face-à-face pour cette étude. Cette focalisation permet de mieux explorer et optimiser la prise de décision et les capacités d'interaction du robot dans un contexte plus contrôlé et direct, tout en posant les bases pour des recherches futures qui pourront réintégrer les interactions en groupe une fois les obstacles techniques surmontés. Cette approche ciblée nous permet d'approfondir dans un premier temps notre compréhension des mécanismes nécessaires pour une interaction proactive humain-robot efficace et naturelle, posant ainsi les fondations pour des avancées significatives dans le domaine de la robotique sociale.

4.5 PERSPECTIVES

Les expériences réalisées ont révélé d'importantes pistes d'amélioration pour le dispositif. Lors des tests à l'hôpital Broca, les variations de scène ont sérieusement compromis le module de prédiction de l'IAB. Le module de suivi du regard s'est avéré inefficace à longue distance, or le regard est une composante essentielle de la prédiction de l'IAB. Il est donc nécessaire de trouver des alternatives plus performantes pour remplacer ce module de vision à longue portée. L'utilisation d'un modèle de langage de grande taille pour analyser le regard à distance pourrait constituer une piste extrêmement prometteuse. En effet, ces modèles de langage possèdent une capacité avancée de traitement et de compréhension contextuelle qui pourrait améliorer l'analyse du regard, même à

longue distance. De plus, ces LLMs pourraient également être en mesure de prédire l'IAB de manière précise et fiable. Pour ce faire, on pourrait définir un prompt détaillant explicitement les critères de l'IAB, accompagné de quelques exemples illustratifs (méthode *few-shots*). Cette approche permettrait de tirer parti de la puissance des nouveaux LLMs pour obtenir des prédictions plus robustes et mieux adaptées aux diverses variations de scène rencontrées.

La mise en œuvre de l'architecture FlowAct nous a conduit à privilégier un système d'engagement basé sur des règles, s'appuyant sur la valeur de l'IAB ou un signal clair de demande d'interaction tel qu'un mouvement de la main. L'apprentissage de ces comportements proactifs d'interaction est une tâche délicate qui nécessite des données pour entraîner un modèle. En l'absence de jeu de données correspondant à cette tâche, l'utilisation d'un large modèle de langage en tant qu'agent décisionnel se révèle particulièrement intéressante. En effet, un LLM, préalablement entraîné sur de vastes corpus de données, possède une compréhension du langage naturel, ce qui lui permet de prendre des décisions basées sur des contextes complexes. Grâce à sa capacité à analyser et à interpréter une variété de signaux linguistiques et comportementaux, il peut identifier les intentions et les besoins des utilisateurs avec précision. Cette aptitude à comprendre le contexte et à générer des réponses appropriées permet de simuler des interactions humaines de manière fluide et naturelle. Cependant, afin de s'intégrer au système FlowAct, un travail linguistique de définition et d'analyse des perceptions est nécessaire en amont pour donner au LLM les meilleures clés de compréhension de la scène, et ainsi équilibrer exhaustivité et optimalité vis-à-vis des objectifs du module de prise de décision.

Ainsi, après avoir exploré un schéma de fonctionnement standard comprenant les étapes de "poste d'observation, engagement, interaction et retour au poste d'observation", représentant la boucle de l'objectif proactif de l'agent dans une salle d'attente d'hôpital, il serait possible d'élargir les situations grâce à un modèle de langage de grande taille. Ce dernier peut continuellement apprendre et s'adapter aux nouveaux comportements et interactions, offrant ainsi une flexibilité et une évolutivité essentielles pour le développement de systèmes interactifs. De plus, il peut intégrer des connaissances provenant de domaines variés, ce qui enrichit les décisions prises et améliore l'expérience utilisateur en rendant les interactions plus pertinentes et personnalisées.

Par ailleurs, l'utilisation du robot Pepper, si elle a permis d'aboutir à une mise en œuvre concrète des contributions de cette thèse, reste délicate. L'intégration de Pepper dans un contexte réaliste, comme un hôpital, reste compliquée, en raison notamment de la performance limitée de ses capteurs. L'ajout d'une caméra, d'un micro et potentiellement d'un lidar sur le Pepper améliorerait considérablement la qualité des interactions. L'agent serait ainsi capable de mieux comprendre les utilisateurs, même s'ils parlent avec une voix faible, de reconnaître facilement ceux avec qui il a déjà interagi et de se déplacer aisément dans son environnement, grâce à un système de SLAM.

En résumé, cette thèse a mis en évidence des pistes de développement pour améliorer la conception de robots multimodaux proactifs d'interaction humain-robot dans des environnements complexes, comme les hôpitaux. Les défis rencontrés ouvrent la voie à des innovations prometteuses, notamment par l'intégration de modèles de langage de grande taille pour améliorer la compréhension et l'interaction avec les utilisateurs. En adoptant des systèmes d'engagement basés sur des règles et en tirant parti des avancées

en intelligence artificielle, il est possible de créer des agents plus réactifs et adaptatifs. Ces améliorations permettront non seulement d'enrichir les interactions, mais aussi de répondre de manière plus précise à l'implémentation de ce concept complexe qu'est la proactivité.

Dans la "Condition de l'homme moderne" (ARENDT 1958), Hannah Arendt écrit que « tout ce que les hommes font ou savent, ou ce dont ils ont l'expérience, a un sens seulement dans la mesure où il est possible d'en parler ». Aujourd'hui, avec les larges modèles de langage, l'humain moderne n'est plus le seul à manipuler la parole. Désormais, il peut la confier aux robots qui, à leur tour, deviennent dépositaires de notre verbe. Comme des échos d'une humanité partagée, ces robots modernes façonnent et propagent nos histoires, nos pensées et nos rêves, prolongeant ainsi notre quête infinie de sens dans un dialogue entre l'humain et la machine. Toutefois, cette vision reste une interprétation philosophique et non une vérité scientifique, qui doit encore donner matière à la réflexion et au débat sur la place et la signification réelle de ces nouvelles technologies vis-à-vis de l'humain.

BIBLIOGRAPHIE

- ARENDE, Hannah (1958). *Condition de l'homme moderne*. Calmann-Lévy.
- ARTHUR, David et Sergei VASSILVITSKII (2007). « k-means++: the advantages of careful seeding ». In : *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '07. New Orleans, Louisiana : Society for Industrial et Applied Mathematics, p. 1027-1035. ISBN : 9780898716245.
- BAEK, Chaehyun, Jung Ju CHOI et Sonya S. KWAK (2014). « Can you touch me? the impact of physical contact on emotional engagement with a robot ». In : *Proceedings of the Second International Conference on Human-Agent Interaction*. HAI 14. Tsukuba, Japan : Association for Computing Machinery, p. 149-152. ISBN : 9781450330350. DOI : [10.1145/2658861.2658909](https://doi.org/10.1145/2658861.2658909). URL : <https://doi.org/10.1145/2658861.2658909>.
- BALTRUŠAITIS, Tadas et al. (2018). « OpenFace 2.0: Facial Behavior Analysis Toolkit ». In : *IEEE International Conference on Automatic Face and Gesture Recognition*.
- BANGOR, Aaron, Philip T. KORTUM et James T. MILLER (2009). « Determining what individual SUS scores mean: adding an adjective rating scale ». In : *Journal of Usability Studies archive* 4, p. 114-123. URL : <https://api.semanticscholar.org/CorpusID:7812093>.
- BARAGLIA, Jimmy et al. (fév. 2017). « Efficient human-robot collaboration: when should a robot take initiative? » In : *The International Journal of Robotics Research*, p. 027836491668825. DOI : [10.1177/0278364916688253](https://doi.org/10.1177/0278364916688253).
- BATEMAN, Thomas S. et J. Michael CRANT (1993). « The proactive component of organizational behavior: A measure and correlates ». In : *Journal of Organizational Behavior* 14.2, p. 103-118. DOI : <https://doi.org/10.1002/job.4030140202>. eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1002/job.4030140202>. URL : <https://onlinelibrary.wiley.com/doi/abs/10.1002/job.4030140202>.
- BAYES, Mr. et Mr. PRICE (1763). « An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. » In : *Philosophical Transactions (1683-1775)* 53, p. 370-418. ISSN : 02607085. URL : <http://www.jstor.org/stable/105741> (visité le 28/05/2024).
- BEN-YOUSSEF, Atef et al. (2017). « UE-HRI: A New Dataset for the Study of User Engagement in Spontaneous Human-Robot Interactions ». In : *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ICMI '17. Glasgow, UK : Association for Computing Machinery, p. 464-472. ISBN : 9781450355438. DOI : [10.1145/3136755.3136814](https://doi.org/10.1145/3136755.3136814). URL : <https://doi.org/10.1145/3136755.3136814>.
- BESL, P.J. et Neil D. MCKAY (1992). « A method for registration of 3-D shapes ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14.2, p. 239-256. DOI : [10.1109/34.121791](https://doi.org/10.1109/34.121791).
- BILLING, E., J. ROSÉN et M. LAMB (2023). « Language Models for Human-Robot Interaction ». In : *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23)*, p. 905-906.

- BISHOP, Christopher (2006). *Pattern Recognition and Machine Learning*. Springer. URL : <https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/>.
- BISHOP, Christopher M. (2007). *Pattern Recognition and Machine Learning*. New York : Springer. ISBN : 978-0-387-31073-2.
- BOCHKOVSKIY, Alexey, Chien-Yao WANG et Hong-Yuan Mark LIAO (2020). *YOLOv4: Optimal Speed and Accuracy of Object Detection*. arXiv : 2004.10934 [cs.CV].
- BOURLARD, Hervé et Nelson MORGAN (1993). « Connectionist Speech Recognition: A Hybrid Approach ». In : URL : <https://api.semanticscholar.org/CorpusID:61058350>.
- BREAZEAL, Cynthia (2003). « Toward sociable robots ». In : *Robotics and Autonomous Systems* 42.3. Socially Interactive Robots, p. 167-175. ISSN : 0921-8890. DOI : [https://doi.org/10.1016/S0921-8890\(02\)00373-1](https://doi.org/10.1016/S0921-8890(02)00373-1). URL : <https://www.sciencedirect.com/science/article/pii/S0921889002003731>.
- (août 2004). *Designing Sociable Robots*. The MIT Press. ISBN : 9780262255837. DOI : [10.7551/mitpress/2376.001.0001](https://doi.org/10.7551/mitpress/2376.001.0001). URL : <https://doi.org/10.7551/mitpress/2376.001.0001>.
- BREMNER, Paul et al. (2019a). « On Proactive, Transparent, and Verifiable Ethical Reasoning for Robots ». In : *Proceedings of the IEEE* 107.3, p. 541-561. DOI : [10.1109/jproc.2019.2898267](https://doi.org/10.1109/jproc.2019.2898267). URL : <https://doi.org/10.1109%2Fjproc.2019.2898267>.
- (2019b). « On Proactive, Transparent, and Verifiable Ethical Reasoning for Robots ». In : *Proceedings of the IEEE* 107.3, p. 541-561. DOI : [10.1109/JPROC.2019.2898267](https://doi.org/10.1109/JPROC.2019.2898267).
- BROOKE, John (nov. 1995). « SUS: A quick and dirty usability scale ». In : *Usability Eval. Ind.* 189.
- BROWN, Tom B. et al. (2020). « Language Models are Few-Shot Learners ». In : *ArXiv abs/2005.14165*.
- BURGOON Judee K. and Buller, David B. (1994). « Interpersonal deception: III. Effects of deceit on perceived communication and nonverbal behavior dynamics ». In : *Journal of Nonverbal Behavior* 18.2, p. 155-184. ISSN : 1573-3653. DOI : [10.1007/BF02170076](https://doi.org/10.1007/BF02170076). URL : <https://doi.org/10.1007/BF02170076>.
- BUYUKGOZ, Sera (mars 2022). « Generating Proactive Communicative Robot Behaviors ». Thèse de doct. Sorbonne Université. URL : <https://theses.hal.science/tel-03863684>.
- C., Ujjwal K. et Jacques CHODOROWSKI (2019). « A Case Study of Adding Proactivity in Indoor Social Robots Using Belief–Desire–Intention (BDI) Model ». In : *Biomimetics* 4.4, p. 74. DOI : [10.3390/biomimetics4040074](https://doi.org/10.3390/biomimetics4040074). URL : <https://doi.org/10.3390%2Fbiomimetics4040074>.
- CARLETTA, Jean et al. (juill. 2005). « The AMI meeting corpus: A pre-announcement ». In : *Lecture Notes in Computer Science*. DOI : [10.1007/11677482_3](https://doi.org/10.1007/11677482_3).
- CARREIRA, Joao et Andrew ZISSERMAN (2018). *Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset*. arXiv : 1705.07750 [cs.CV].
- CASTALDO, Fabio et al. (2008). « Stream-based speaker segmentation using speaker factors and eigenvoices ». In : *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, p. 4133-4136. DOI : [10.1109/ICASSP.2008.4518564](https://doi.org/10.1109/ICASSP.2008.4518564).

- CHEN, Chin S., Chia J. LIN et Chun C. LAI (2022). « Non-Contact Service Robot Development in Fast-Food Restaurants ». In : *IEEE Access* 10, p. 31466-31479. DOI : [10.1109/ACCESS.2022.3155661](https://doi.org/10.1109/ACCESS.2022.3155661).
- CHEN, Xiaohui et al. (2024). *Does ChatGPT and Whisper Make Humanoid Robots More Reliable?* arXiv : [2402.07095 \[cs.R0\]](https://arxiv.org/abs/2402.07095). URL : <https://arxiv.org/abs/2402.07095>.
- CHEN, Zhe et al. (2024). *InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks*. arXiv : [2312.14238 \[cs.CV\]](https://arxiv.org/abs/2312.14238).
- CHIANG, Wei-Lin et al. (2023). *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*. URL : <https://lmsys.org/blog/2023-03-30-vicuna/>.
- CHITTA, Sachin et al. (2017). « ros_control: A generic and simple control framework for ROS ». In : *Journal of Open Source Software* 2.20, p. 456. DOI : [10.21105/joss.00456](https://doi.org/10.21105/joss.00456). URL : <https://doi.org/10.21105/joss.00456>.
- CHO, Kyunghyun et al. (2014). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. arXiv : [1406.1078 \[cs.CL\]](https://arxiv.org/abs/1406.1078). URL : <https://arxiv.org/abs/1406.1078>.
- CHOROWSKI, Jan et al. (2015). *Attention-Based Models for Speech Recognition*. arXiv : [1506.07503 \[cs.CL\]](https://arxiv.org/abs/1506.07503). URL : <https://arxiv.org/abs/1506.07503>.
- CHUNG, Joon Son, Bong-Jin LEE et Icksang HAN (2019). *Who said that?: Audio-visual speaker diarisation of real-world meetings*. DOI : [10.48550/ARXIV.1906.10042](https://arxiv.org/abs/1906.10042). URL : <https://arxiv.org/abs/1906.10042>.
- CLADIÈRE, Tristan et al. (2023). « BENet: A Lightweight Bottom-Up Framework for Context-Aware Emotion Recognition ». In : *Advanced Concepts for Intelligent Vision Systems: 21st International Conference, ACIVS 2023 Kumamoto, Japan, August 21–23, 2023 Proceedings*. Kumamoto, Japan : Springer-Verlag, p. 100-111. ISBN : 978-3-031-45381-6.
- COLEMAN, David et al. (2014). *Reducing the Barrier to Entry of Complex Robotic Software: a MoveIt! Case Study*. arXiv : [1404.3785 \[cs.R0\]](https://arxiv.org/abs/1404.3785).
- COOPER, Sara et al. (août 2020). « ARI: the Social Assistive Robot and Companion ». In : p. 745-751. DOI : [10.1109/R0-MAN47096.2020.9223470](https://doi.org/10.1109/R0-MAN47096.2020.9223470).
- DAI, Jifeng et al. (2016). *R-FCN: Object Detection via Region-based Fully Convolutional Networks*. arXiv : [1605.06409 \[cs.CV\]](https://arxiv.org/abs/1605.06409).
- DALAL, N. et B. TRIGGS (2005). « Histograms of oriented gradients for human detection ». In : *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. T. 1, 886-893 vol. 1. DOI : [10.1109/CVPR.2005.177](https://doi.org/10.1109/CVPR.2005.177).
- DAS, Dipankar et al. (2015). « Supporting Human-Robot Interaction Based on the Level of Visual Focus of Attention ». In : *IEEE Transactions on Human-Machine Systems* 45.6, p. 664-675. DOI : [10.1109/THMS.2015.2445856](https://doi.org/10.1109/THMS.2015.2445856).
- DAUTENHAHN, Kerstin (1998). « THE ART OF DESIGNING SOCIALLY INTELLIGENT AGENTS: SCIENCE, FICTION, AND THE HUMAN IN THE LOOP ». In : *Applied Artificial Intelligence* 12.7-8, p. 573-617. DOI : [10.1080/088395198117550](https://doi.org/10.1080/088395198117550). eprint : <https://doi.org/10.1080/088395198117550>. URL : <https://doi.org/10.1080/088395198117550>.
- (2007). « Socially intelligent robots: dimensions of human-robot interaction ». English. In : *Philosophical Transactions of the Royal Society B: Biological Sciences* 362.1480, p. 679-704. DOI : <https://doi.org/10.1098/rstb.2006.2004>.

- DEMPSTER, A. P., N. M. LAIRD et D. B. RUBIN (déc. 1977). « Maximum Likelihood from Incomplete Data Via the EM Algorithm ». In : *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, p. 1-22. ISSN : 0035-9246. DOI : [10.1111/j.2517-6161.1977.tb01600.x](https://doi.org/10.1111/j.2517-6161.1977.tb01600.x). URL : <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>.
- DEVLIN, Jacob et al. (2019). « BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding ». In : *North American Chapter of the Association for Computational Linguistics (NAACL)*.
- DHAUSSY, Timothée, Bassam JABAÏAN et Fabrice LEFÈVRE (2023). « Interaction Acceptance Modelling and Estimation for a Proactive Engagement in the Context of Human-Robot Interactions ». In : *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, p. 3069-3074.
- DHAUSSY, Timothée, Bassam JABAÏAN, Fabrice LEFÈVRE et Radu HORAUD (2023). « Audio-Visual Speaker Diarization in the Framework of Multi-User Human-Robot Interaction ». In : *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 1-5. DOI : [10.1109/ICASSP49357.2023.10096295](https://doi.org/10.1109/ICASSP49357.2023.10096295).
- DIXIT, Pranav (2013). « Robotic Trash Cans Follow You Around Until You Pick Up Your Trash ». In : *Fast Company*. URL : <https://www.fastcompany.com/3021754/robotic-trash-cans-follow-you-around-until-you-pick-up-your-trash>.
- DIXIT, Priyanka et Sanjay SILAKARI (2021). « Deep Learning Algorithms for Cybersecurity Applications: A Technological and Status Review ». In : *Computer Science Review* 39, p. 100317. ISSN : 1574-0137. DOI : <https://doi.org/10.1016/j.cosrev.2020.100317>. URL : <https://www.sciencedirect.com/science/article/pii/S1574013720304172>.
- DIXON, Philip M. (2006). « Nearest Neighbor Methods ». In : *Encyclopedia of Environmetrics*. John Wiley & Sons, Ltd. ISBN : 9780470057339. DOI : <https://doi.org/10.1002/9780470057339.van007>.
- ELICES, J. (2020). *Medieval Robots? They Were Just One of This Muslim Inventor's Creations*. Accessed: 2023-11-29. URL : <https://www.nationalgeographic.co.uk/history-and-civilisation/2020/08/medieval-robots-they-were-just-one-of-this-muslim-inventors>.
- EYBEN, F. et al. (2016). « The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing ». In : *IEEE Transactions on Affective Computing* 7.02, p. 190-202. ISSN : 1949-3045. DOI : [10.1109/TAFFC.2015.2457417](https://doi.org/10.1109/TAFFC.2015.2457417).
- EYBEN, Florian, Martin WÖLLMER et Björn SCHULLER (2010). « openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor ». In : *Proc. ACM Multimedia (MM)*. ACM. Florence, Italy, p. 1459-1462. ISBN : 978-1-60558-933-6.
- FAOUZI, Johann et O. COLLIOT (2023). « Classic machine learning methods ». In : *ArXiv* abs/2310.11470. DOI : [10.48550/arXiv.2310.11470](https://doi.org/10.48550/arXiv.2310.11470).
- FAWCETT, Tom (2006). « An introduction to ROC analysis ». In : *Pattern Recognition Letters* 27.8. ROC Analysis in Pattern Recognition, p. 861-874. ISSN : 0167-8655. DOI : <https://doi.org/10.1016/j.patrec.2005.10.010>. URL : <https://www.sciencedirect.com/science/article/pii/S016786550500303X>.
- FELZENSZWALB, Pedro, David MCALLESTER et Deva RAMANAN (2008). « A discriminatively trained, multiscale, deformable part model ». In : *2008 IEEE Conference on Computer Vision and Pattern Recognition*, p. 1-8. DOI : [10.1109/CVPR.2008.4587597](https://doi.org/10.1109/CVPR.2008.4587597).

- FINSTAD, Kraig (sept. 2010). « The Usability Metric for User Experience ». In : *Interacting with Computers* 22, p. 323-327. DOI : [10.1016/j.intcom.2010.04.004](https://doi.org/10.1016/j.intcom.2010.04.004).
- FISCHER, Tobias, Hyung Jin CHANG et Yiannis DEMIRIS (2018). « RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments ». In : *Proceedings of the European Conference on Computer Vision (ECCV)*.
- FONG, Terrence, Illah NOURBAKSH et Kerstin DAUTENHAHN (2003). « A survey of socially interactive robots ». In : *Robotics and Autonomous Systems* 42.3. Socially Interactive Robots, p. 143-166. ISSN : 0921-8890. DOI : [https://doi.org/10.1016/S0921-8890\(02\)00372-X](https://doi.org/10.1016/S0921-8890(02)00372-X). URL : <https://www.sciencedirect.com/science/article/pii/S092188900200372X>.
- FONTAINE, Vincent, Christophe RIS et Jean-Marc BOITE (1997). « Nonlinear discriminant analysis for improved speech recognition ». In : *EUROSPEECH*. URL : <https://api.semanticscholar.org/CorpusID:14666644>.
- FORBRIG, Peter, Alexandru BUNDEA et Sebastian BADER (juin 2021). « Engineering the Interaction of a Humanoid Robot Pepper with Post-Stroke Patients During Training Tasks ». In : p. 38-43. DOI : [10.1145/3459926.3464756](https://doi.org/10.1145/3459926.3464756).
- GAO, Tian et al. (2018). « Densely Connected Progressive Learning for LSTM-Based Speech Enhancement ». In : *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5054-5058. DOI : [10.1109/ICASSP.2018.8461861](https://doi.org/10.1109/ICASSP.2018.8461861).
- GARDECKI, Arkadiusz et al. (2018). « The Pepper Humanoid Robot in Front Desk Application ». In : *2018 Progress in Applied Electrical Engineering (PAEE)*, p. 1-7. DOI : [10.1109/PAEE.2018.8441069](https://doi.org/10.1109/PAEE.2018.8441069).
- GEBRU, Israel D., Sileye BA et al. (2018). « Audio-Visual Speaker Diarization Based on Spatiotemporal Bayesian Fusion ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.5, p. 1086-1099. DOI : [10.1109/tpami.2017.2648793](https://doi.org/10.1109/tpami.2017.2648793). URL : <https://doi.org/10.1109/2Ftpami.2017.2648793>.
- GEBRU, Israel D., Silèye BA et al. (2015). « Audio-Visual Speech-Turn Detection and Tracking ». In : *Proceedings of the 12th International Conference on Latent Variable Analysis and Signal Separation - Volume 9237*. LVA/ICA 2015. Liberec, Czech Republic : Springer-Verlag, p. 143-151. ISBN : 9783319224817.
- GIRSHICK, R. et al. (2014). « Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation ». In : *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA : IEEE Computer Society, p. 580-587. DOI : [10.1109/CVPR.2014.81](https://doi.org/10.1109/CVPR.2014.81). URL : <https://doi.ieeecomputersociety.org/10.1109/CVPR.2014.81>.
- GIRSHICK, Ross (2015). *Fast R-CNN*. arXiv : [1504.08083](https://arxiv.org/abs/1504.08083) [cs.CV].
- GOLAM RASHED, MD (2016). « Observing People's Behaviors in Public Spaces for Initiating Proactive Human-Robot Interaction by Social Robots ». Ph.D. Dissertation. Japan : Saitama University. URL : <https://ci.nii.ac.jp/naid/500001036626>.
- GONZÁLEZ-JIMÉNEZ, J., C. GALINDO et J.R. RUIZ-SARMIENTO (2012). « Technical improvements of the Giraff telepresence robot based on users' evaluation ». In : *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, p. 827-832. DOI : [10.1109/ROMAN.2012.6343854](https://doi.org/10.1109/ROMAN.2012.6343854).
- GOODFELLOW, Ian J. et al. (2020). « Generative Adversarial Networks ». In : *Communications of the ACM* 63, p. 139-144. DOI : [10.1145/3422622](https://doi.org/10.1145/3422622).

- GRAEFE, Volker et Rainer BISCHOFF (2009). « From ancient machines to intelligent robots — A technical evolution — ». In : *2009 9th International Conference on Electronic Measurement & Instruments*, p. 3-418-3-431. DOI : [10.1109/ICEMI.2009.5274297](https://doi.org/10.1109/ICEMI.2009.5274297).
- GRANT, Adam M. et Susan J. ASHFORD (2008). « The dynamics of proactivity at work ». In : *Research in Organizational Behavior* 28, p. 3-34. ISSN : 0191-3085. DOI : <https://doi.org/10.1016/j.riob.2008.04.002>. URL : <https://www.sciencedirect.com/science/article/pii/S0191308508000038>.
- GRAVES, Alex (2012). *Sequence Transduction with Recurrent Neural Networks*. arXiv : [1211.3711](https://arxiv.org/abs/1211.3711) [cs.NE]. URL : <https://arxiv.org/abs/1211.3711>.
- GRAVES, Alex, Santiago FERNÁNDEZ et al. (2006). « Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks ». In : *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA : Association for Computing Machinery, p. 369-376. ISBN : 1595933832. DOI : [10.1145/1143844.1143891](https://doi.org/10.1145/1143844.1143891). URL : <https://doi.org/10.1145/1143844.1143891>.
- GRAVES, Alex et Navdeep JAITLY (2014). « Towards end-to-end speech recognition with recurrent neural networks ». In : *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. ICML'14. Beijing, China : JMLR.org, II-1764-II-1772.
- GRAVES, Alex, Greg WAYNE et Ivo DANIHELKA (2014). *Neural Turing Machines*. arXiv : [1410.5401](https://arxiv.org/abs/1410.5401) [cs.NE]. URL : <https://arxiv.org/abs/1410.5401>.
- GRONDIN, Francois et Francois MICHAUD (2018). *Lightweight and Optimized Sound Source Localization and Tracking Methods for Open and Closed Microphone Array Configurations*. DOI : [10.48550/ARXIV.1812.00115](https://arxiv.org/abs/1812.00115). URL : <https://arxiv.org/abs/1812.00115>.
- GROSINGER, Jasmin (2022). « On Proactive Human-AI Systems ». In : *International Workshop on Artificial Intelligence and Cognition*. (Visité le 15/12/2022).
- GROSINGER, Jasmin, Federico PECORA et Alessandro SAFFIOTTI (jan. 2016). « Making Robots Proactive through Equilibrium Maintenance ». In : — (2017). « Proactivity through equilibrium maintenance with fuzzy desirability ». In : *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, p. 2117-2122. DOI : [10.1109/SMC.2017.8122932](https://doi.org/10.1109/SMC.2017.8122932).
- (mai 2018). « Robots that Maintain Equilibrium: Proactivity by Reasoning About User Intentions and Preferences ». In : *Pattern Recognition Letters* 118. DOI : [10.1016/j.patrec.2018.05.014](https://doi.org/10.1016/j.patrec.2018.05.014).
- GUGGEMOS, Josef, Sabine SEUFERT et Stefan SONDEREGGER (2020). « Humanoid robots in higher education: Evaluating the acceptance of Pepper in the context of an academic writing course using the UTAUT ». In : *Br. J. Educ. Technol.* 51, p. 1864-1883. URL : <https://api.semanticscholar.org/CorpusID:221766701>.
- GUPTA, Abhay et al. (2022). *DAiSEE: Towards User Engagement Recognition in the Wild*. arXiv : [1609.01885](https://arxiv.org/abs/1609.01885) [cs.CV].
- HAN, Jeonghye et al. (2009). « The birth story of robot, IROBIQ for children's tolerance ». In : *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, p. 318-318. DOI : [10.1109/ROMAN.2009.5326249](https://doi.org/10.1109/ROMAN.2009.5326249).
- HAN, Yean et al. (2018). « A human-robot interface to improve facial expression recognition in subjects with Autism Spectrum Disorder ». In : *2018 9th International Conference*

- on *Awareness Science and Technology (iCAST)*, p. 179-184. DOI : [10.1109/ICAWST.2018.8517228](https://doi.org/10.1109/ICAWST.2018.8517228).
- HE, Kaiming, Georgia GKIOXARI et al. (2018). *Mask R-CNN*. arXiv : [1703.06870 \[cs.CV\]](https://arxiv.org/abs/1703.06870).
- HE, Kaiming, Xiangyu ZHANG et al. (2014). « Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition ». In : *Lecture Notes in Computer Science*. Springer International Publishing, p. 346-361. ISBN : 9783319105789.
- HECHT-NIELSEN, R. (1989). « Theory of the backpropagation neural network ». In : *International 1989 Joint Conference on Neural Networks*, 593-605 vol.1. DOI : [10.1016/0893-6080\(88\)90469-8](https://doi.org/10.1016/0893-6080(88)90469-8).
- HERSHEY, John et Javier MOVELLAN (1999). « Audio Vision: Using Audio-Visual Synchrony to Locate Sounds ». In : *Advances in Neural Information Processing Systems*. Sous la dir. de S. SOLLA, T. LEEN et K. MÜLLER. T. 12. MIT Press. URL : <https://proceedings.neurips.cc/paper/1999/file/b618c3210e934362ac261db280128c22-Paper.pdf>.
- HOCHREITER, Sepp et Jürgen SCHMIDHUBER (1997). « Long Short-Term Memory ». In : *Neural Computation* 9, p. 1735-1780. DOI : [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735).
- HOUGH, Julian et David SCHLANGEN (2017). « It's Not What You Do, It's How You Do It: Grounding Uncertainty for a Simple Robot ». In : *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '17. Vienna, Austria : Association for Computing Machinery, p. 274-282. ISBN : 9781450343367. DOI : [10.1145/2909824.3020214](https://doi.org/10.1145/2909824.3020214). URL : <https://doi.org/10.1145/2909824.3020214>.
- HURST, Nikki et al. (2020). « Social and Emotional Skills Training with Embodied Moxie ». In : *ArXiv* abs/2004.12962. URL : <https://api.semanticscholar.org/CorpusID:216553473>.
- INOUE, Koji et al. (2018). « Engagement recognition by a latent character model based on multimodal listener behaviors in spoken dialogue ». In : *APSIPA Transactions on Signal and Information Processing* 7, e9. DOI : [10.1017/ATSIP.2018.11](https://doi.org/10.1017/ATSIP.2018.11).
- JAN, Bilal et al. (2019). « Deep learning in big data Analytics: A comparative study ». In : *Computers & Electrical Engineering* 75, p. 275-287. ISSN : 0045-7906. DOI : <https://doi.org/10.1016/j.compeleceng.2017.12.009>. URL : <https://www.sciencedirect.com/science/article/pii/S0045790617315835>.
- JANISZEK, David et al. (oct. 2011). « De l'usage de Nao (robot humanoïde) dans l'apprentissage de l'informatique ». In : URL : <https://www.epi.asso.fr/revue/articles/a1206f.htm>.
- KALASHNIKOVA, Natalia et al. (oct. 2023). « Do We Speak to Robots Looking Like Humans As We Speak to Humans? A Study of Pitch in French Human-Machine and Human-Human Interactions ». In : p. 141-145. DOI : [10.1145/3610661.3617990](https://doi.org/10.1145/3610661.3617990).
- KALMAN, Rudolph Emil (1960). « A New Approach to Linear Filtering and Prediction Problems ». In : *Transactions of the ASME—Journal of Basic Engineering* 82.Series D, p. 35-45.
- KANDA, Takayuki, Dylan GLAS et al. (jan. 2010). « Abstracting People's Trajectories for Social Robots to Proactively Approach Customers ». In : *Robotics, IEEE Transactions on* 25, p. 1382-1396. DOI : [10.1109/TR0.2009.2032969](https://doi.org/10.1109/TR0.2009.2032969).

- KANDA, Takayuki, Masahiro SHIOMI et al. (2009). « An affective guide robot in a shopping mall ». In : *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, p. 173-180. DOI : [10.1145/1514095.1514127](https://doi.org/10.1145/1514095.1514127).
- KAPLAN, Frédéric (2000). « Talking AIBO: First Experimentation of Verbal Interactions with an Autonomous Four-legged Robot ». In : *Learning to Behave: Interacting agents CELE-TWENTE Workshop on Language Technology*. Sony Computer Science Laboratories, Paris, p. 57-63.
- KATO, Yusuke, Takayuki KANDA et Hiroshi ISHIGURO (mars 2015). « May I help you?: Design of Human-like Polite Approaching Behavior ». In : t. 2015, p. 35-42. DOI : [10.1145/2696454.2696463](https://doi.org/10.1145/2696454.2696463).
- KENNY, Patrick, Douglas REYNOLDS et Fabio CASTALDO (2010). « Diarization of Telephone Conversations Using Factor Analysis ». In : *IEEE Journal of Selected Topics in Signal Processing* 4.6, p. 1059-1070. DOI : [10.1109/JSTSP.2010.2081790](https://doi.org/10.1109/JSTSP.2010.2081790).
- EL-KHOURY, Elie, Christine SENAC et Julien PINQUIER (2009). « Improved speaker diarization system for meetings ». In : *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP '09*. USA : IEEE Computer Society, p. 4097-4100. ISBN : 9781424423538. DOI : [10.1109/ICASSP.2009.4960529](https://doi.org/10.1109/ICASSP.2009.4960529). URL : <https://doi.org/10.1109/ICASSP.2009.4960529>.
- KLOTZ, David et al. (2011). « Engagement-based Multi-party Dialog with a Humanoid Robot ». In : *Proceedings of the SIGDIAL 2011 Conference*. Sous la dir. de Joyce Y. CHAI et al. Portland, Oregon : Association for Computational Linguistics, p. 341-343. URL : <https://aclanthology.org/W11-2042>.
- KNAPP, Mark L et Judith A HALL (1972). *Nonverbal Communication in Human Interaction*. Holt, Rinehart et Winston.
- KOIDE, Y. et al. (jan. 2004). « An approach to integrating an interactive guide robot with ubiquitous sensors ». In : 2500-2505 vol.3. ISBN : 0-7803-8463-6. DOI : [10.1109/IROS.2004.1389784](https://doi.org/10.1109/IROS.2004.1389784).
- KRIMAN, Samuel et al. (2019). *QuartzNet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions*. arXiv : [1910.10261 \[eess.AS\]](https://arxiv.org/abs/1910.10261). URL : <https://arxiv.org/abs/1910.10261>.
- KRIZHEVSKY, Alex, Ilya SUTSKEVER et Geoffrey E HINTON (2012). « ImageNet classification with deep convolutional neural networks ». In : *Advances in neural information processing systems*, p. 1097-1105.
- KUNO, Yoshinori et al. (2007). « Museum guide robot based on sociological interaction analysis ». In : *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '07*. San Jose, California, USA : Association for Computing Machinery, p. 1191-1194. ISBN : 9781595935939. DOI : [10.1145/1240624.1240804](https://doi.org/10.1145/1240624.1240804). URL : <https://doi.org/10.1145/1240624.1240804>.
- LANDINI, Federico et al. (2020). *Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks*. DOI : [10.48550/ARXIV.2012.14952](https://arxiv.org/abs/2012.14952). URL : <https://arxiv.org/abs/2012.14952>.
- LECUN, Y. et al. (1998). « Gradient-based learning applied to document recognition ». In : *Proceedings of the IEEE* 86.11, p. 2278-2324. DOI : [10.1109/5.726791](https://doi.org/10.1109/5.726791).

- LEE, Hyojun et al. (2021). « A Robot Operating System Framework for Secure UAV Communications ». In : *Sensors* 21.4. ISSN : 1424-8220. DOI : [10.3390/s21041369](https://doi.org/10.3390/s21041369). URL : <https://www.mdpi.com/1424-8220/21/4/1369>.
- LEMAIGNAN, Séverin, Raquel Ros et al. (oct. 2010). « ORO, a knowledge management platform for cognitive architectures in robotics ». In : p. 3548-3553. DOI : [10.1109/IRoS.2010.5649547](https://doi.org/10.1109/IRoS.2010.5649547).
- LEMAIGNAN, Séverin, Mathieu WARNIER et al. (2017). « Artificial cognition for social human-robot interaction: An implementation ». In : *Artificial Intelligence* 247. Special Issue on AI and Robotics, p. 45-69. ISSN : 0004-3702. DOI : <https://doi.org/10.1016/j.artint.2016.07.002>. URL : <https://www.sciencedirect.com/science/article/pii/S0004370216300790>.
- LI, Jason et al. (2019). « Jasper: An end-to-end convolutional neural acoustic model ». In : *arXiv preprint arXiv:1904.03288*.
- LI, Shan et Weihong DENG (juill. 2022). « Deep Facial Expression Recognition: A Survey ». In : *IEEE Transactions on Affective Computing* 13.3, p. 1195-1215. ISSN : 2371-9850. DOI : [10.1109/taffc.2020.2981446](https://doi.org/10.1109/taffc.2020.2981446). URL : <http://dx.doi.org/10.1109/TAFFC.2020.2981446>.
- LI, Yue, Koen HINDRIKS et Florian KUNNEMAN (2024). « Single-Channel Robot Ego-Speech Filtering during Human-Robot Interaction ». In : *2024 International Symposium on Technological Advances in Human-Robot Interaction (TAHRI 2024)*. Boulder, CO, USA. ISBN : 979-8-4007-1661-4. DOI : [10.1145/3648536.3648539](https://doi.org/10.1145/3648536.3648539).
- LIANG, Percy et al. (juill. 2006). « An End-to-End Discriminative Approach to Machine Translation ». In : *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sous la dir. de Nicoletta CALZOLARI, Claire CARDIE et Pierre ISABELLE. Sydney, Australia : Association for Computational Linguistics, p. 761-768. DOI : [10.3115/1220175.1220271](https://doi.org/10.3115/1220175.1220271). URL : <https://aclanthology.org/P06-1096>.
- LIN, Tsung-Yi, Piotr DOLLÁR et al. (2017). *Feature Pyramid Networks for Object Detection*. arXiv : [1612.03144 \[cs.CV\]](https://arxiv.org/abs/1612.03144).
- LIN, Tsung-Yi, Priya GOYAL et al. (2018). *Focal Loss for Dense Object Detection*. arXiv : [1708.02002 \[cs.CV\]](https://arxiv.org/abs/1708.02002).
- LIU, Haotian et al. (2023). *Visual Instruction Tuning*. arXiv : [2304.08485 \[cs.CV\]](https://arxiv.org/abs/2304.08485).
- LIU, Wei et al. (2016). « SSD: Single Shot MultiBox Detector ». In : *Lecture Notes in Computer Science*. Springer International Publishing, p. 21-37. ISBN : 9783319464480.
- MACENSKI, Steven et al. (mai 2022). « Robot Operating System 2: Design, architecture, and uses in the wild ». In : *Science Robotics* 7.66. ISSN : 2470-9476. DOI : [10.1126/scirobotics.abm6074](https://doi.org/10.1126/scirobotics.abm6074). URL : <http://dx.doi.org/10.1126/scirobotics.abm6074>.
- MAHDI, Hamza et al. (2020). « Creating MyJay: A New Design for Robot-Assisted Play for Children with Physical Special Needs ». In : *Social Robotics*. Sous la dir. d'Alan R. WAGNER et al. Cham : Springer International Publishing, p. 676-687. ISBN : 978-3-030-62056-1.
- MCCANN, Hugh J. et M. E. BRATMAN (1991). « Intention, Plans, and Practical Reason. » In : *Noûs* 25.2, p. 230. DOI : [10.2307/2215590](https://doi.org/10.2307/2215590). URL : <https://doi.org/10.2307/2215590>.

- MCLACHLAN, G. (juin 1999). « Mahalanobis Distance ». In : *Resonance* 4, p. 20-26. DOI : [10.1007/BF02834632](https://doi.org/10.1007/BF02834632).
- MEHLMANN, Gregor et al. (2014). « Exploring a Model of Gaze for Grounding in Multimodal HRI ». In : *Proceedings of the 16th International Conference on Multimodal Interaction. ICMI '14*. Istanbul, Turkey : Association for Computing Machinery, p. 247-254. ISBN : 9781450328852. DOI : [10.1145/2663204.2663275](https://doi.org/10.1145/2663204.2663275). URL : <https://doi.org/10.1145/2663204.2663275>.
- MORI, Masahiro, Karl F. MACDORMAN et Norri KAGEKI (2012). « The Uncanny Valley [From the Field] ». In : *IEEE Robotics & Automation Magazine* 19.2, p. 98-100. DOI : [10.1109/MRA.2012.2192811](https://doi.org/10.1109/MRA.2012.2192811).
- MOULIN-FRIER, Clément et al. (déc. 2018). « DAC-h3: A Proactive Robot Cognitive Architecture to Acquire and Express Knowledge About the World and the Self ». In : *IEEE Transactions on Cognitive and Developmental Systems* 10, p. 1005-1022. DOI : [10.1109/TCDS.2017.2754143](https://doi.org/10.1109/TCDS.2017.2754143).
- MUNAWAR, Asim et al. (2018). « MaestROB: A Robotics Framework for Integrated Orchestration of Low-Level Control and High-Level Reasoning ». In : *2018 IEEE International Conference on Robotics and Automation (ICRA)*, p. 527-534. DOI : [10.1109/ICRA.2018.8462870](https://doi.org/10.1109/ICRA.2018.8462870).
- NAKANO, Yukiko et Ryo ISHII (fév. 2010). « Estimating user's engagement from eye-gaze behaviors in human-agent conversations ». In : p. 139-148. DOI : [10.1145/1719970.1719990](https://doi.org/10.1145/1719970.1719990).
- NJIFENJOU, Ahmed et al. (2024). *Role-Play Zero-Shot Prompting with Large Language Models for Open-Domain Human-Machine Conversation*. arXiv : [2406.18460](https://arxiv.org/abs/2406.18460) [cs.CL]. URL : <https://arxiv.org/abs/2406.18460>.
- NOGUCHI, Yohei, Hiroko KAMIDE et Fumihide TANAKA (2020). « Personality Traits for a Social Mediator Robot Encouraging Elderly Self-Disclosure on Loss Experiences ». In : *J. Hum.-Robot Interact.* 9.3. DOI : [10.1145/3377342](https://doi.org/10.1145/3377342). URL : <https://doi.org/10.1145/3377342>.
- PAPADOPOULOS, Fotios et al. (2016). « Do relative positions and proxemics affect the engagement in a Human-Robot collaborative scenario? ». In : *Interaction Studies* 17, p. 321-347. URL : <https://api.semanticscholar.org/CorpusID:151914961>.
- PARK, Yoon Jung et al. (2018). « ChildAR-Bot: Educational Playing Projection-Based AR Robot for Children ». In : *Proceedings of the 26th ACM International Conference on Multimedia. MM '18*. Seoul, Republic of Korea : Association for Computing Machinery, p. 1278-1282. ISBN : 9781450356657. DOI : [10.1145/3240508.3241362](https://doi.org/10.1145/3240508.3241362). URL : <https://doi.org/10.1145/3240508.3241362>.
- PARKER, Sharon K., Uta K. BINDL et Karoline STRAUSS (2010). « Making Things Happen: A Model of Proactive Motivation ». In : *Journal of Management* 36.4, p. 827-856. DOI : [10.1177/0149206310363732](https://doi.org/10.1177/0149206310363732). eprint : <https://doi.org/10.1177/0149206310363732>. URL : <https://doi.org/10.1177/0149206310363732>.
- PASSLER BATES, Danika et James E. YOUNG (2020). « SnuggleBot: A Novel Cuddly Companion Robot Design ». In : *Proceedings of the 8th International Conference on Human-Agent Interaction. HAI '20*. Virtual Event, USA : Association for Computing Machinery, p. 260-262. ISBN : 9781450380546. DOI : [10.1145/3406499.3418772](https://doi.org/10.1145/3406499.3418772). URL : <https://doi.org/10.1145/3406499.3418772>.

- PATACCHIOLA, Massimiliano et Angelo CANGELOSI (2017). « Head pose estimation in the wild using Convolutional Neural Networks and adaptive gradient methods ». In : *Pattern Recognition* 71, p. 132-143. ISSN : 0031-3203. DOI : <https://doi.org/10.1016/j.patcog.2017.06.009>. URL : <https://www.sciencedirect.com/science/article/pii/S0031320317302327>.
- PENG, Zhenhui et al. (2019). « Design and Evaluation of Service Robot's Proactivity in Decision-Making Support Process ». In : *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. DOI : [10.1145/3290605.3300328](https://doi.org/10.1145/3290605.3300328). URL : <https://doi.org/10.1145/3290605.3300328>.
- PERERA, Vittorio et al. (2017). *Setting Up Pepper For Autonomous Navigation And Personalized Interaction With Users*. arXiv : [1704.04797](https://arxiv.org/abs/1704.04797) [cs.R0].
- PIASEK, Joanna et Katarzyna WIECZOROWSKA-TOBIS (2018). « Acceptance and Long-Term Use of a Social Robot by Elderly Users in a Domestic Environment ». In : *2018 11th International Conference on Human System Interaction (HSI)*, p. 478-482. DOI : [10.1109/HSI.2018.8431348](https://doi.org/10.1109/HSI.2018.8431348).
- PIERSON, Harry A. et Michael S. GASHLER (2017). *Deep Learning in Robotics: A Review of Recent Research*. arXiv : [1707.07217](https://arxiv.org/abs/1707.07217) [cs.R0]. URL : <https://arxiv.org/abs/1707.07217>.
- PRESCOTT, Tony et al. (mars 2019a). « Memory and mental time travel in humans and social robots ». In : *Philosophical Transactions B: Biological Sciences* 374. DOI : [10.1098/rstb.2018.0025](https://doi.org/10.1098/rstb.2018.0025).
- (mars 2019b). « Memory and mental time travel in humans and social robots ». In : *Philosophical Transactions B: Biological Sciences* 374. DOI : [10.1098/rstb.2018.0025](https://doi.org/10.1098/rstb.2018.0025).
- QIAN, Xinyuan et al. (2019). « Multi-Speaker Tracking From an Audio-Visual Sensing Device ». In : *IEEE Transactions on Multimedia* 21.10, p. 2576-2588. DOI : [10.1109/TMM.2019.2902489](https://doi.org/10.1109/TMM.2019.2902489).
- QUIGLEY, Morgan et al. (2009). « ROS: an open-source Robot Operating System ». In : *ICRA workshop on open source software*. T. 3. 3.2. Kobe, Japan, p. 5.
- RADFORD, Alec et al. (2022). « Robust Speech Recognition via Large-Scale Weak Supervision ». In : *arXiv preprint arXiv:2212.04356*.
- RAJENDRAN, Gayathri, Uma V et Bettina O'BRIEN (2022). « Unified robot task and motion planning with extended planner using ROS simulator ». In : *Journal of King Saud University - Computer and Information Sciences* 34.9, p. 7468-7481. ISSN : 1319-1578. DOI : <https://doi.org/10.1016/j.jksuci.2021.07.002>. URL : <https://www.sciencedirect.com/science/article/pii/S1319157821001695>.
- RANE, Pranav, Varun MHATRE et Lakshmi KURUP (2014). « Study of a Home Robot: JIBO ». In : *International Journal of Engineering Research & Technology (IJERT)* 3.10. This work is licensed under a Creative Commons Attribution 4.0 International License., ISSN: 2278-0181. URL : www.ijert.org/IJERTV3IS100361.
- RASHED, Md. Golam et al. (nov. 2018). « A Study on Proactive Methods for Initiating Interaction with Human by Social Robots ». In :
- REDMON, Joseph et al. (2016). *You Only Look Once: Unified, Real-Time Object Detection*. arXiv : [1506.02640](https://arxiv.org/abs/1506.02640) [cs.CV].
- REYNOLDS, D.A. et P. TORRES-CARRASQUILLO (2005). « Approaches and applications of audio diarization ». In : *Proceedings. (ICASSP '05). IEEE International Conference on*

- Acoustics, Speech, and Signal Processing, 2005*. T. 5, v/953-v/956 Vol. 5. DOI : [10.1109/ICASSP.2005.1416463](https://doi.org/10.1109/ICASSP.2005.1416463).
- « Robotic cat is the perfect pet for seniors » (2015). In : *Business Insider*. URL : <https://www.businessinsider.com/robotic-cat-is-the-perfect-pet-for-seniors-2015-11?r=US%5C&IR=T>.
- ROLLER, Stephen et al. (2020). *Recipes for building an open-domain chatbot*. arXiv : [2004.13637](https://arxiv.org/abs/2004.13637) [cs.CL]. URL : <https://arxiv.org/abs/2004.13637>.
- ROMERO-GARCÉS, Adrián et al. (2015). « Testing a Fully Autonomous Robotic Salesman in Real Scenarios ». In : *2015 IEEE International Conference on Autonomous Robot Systems and Competitions*, p. 124-130. DOI : [10.1109/ICARSC.2015.19](https://doi.org/10.1109/ICARSC.2015.19).
- ROSSI, Alessandra et al. (2022). *A ROS Architecture for Personalised HRI with a Bartender Social Robot*. arXiv : [2203.06631](https://arxiv.org/abs/2203.06631) [cs.R0]. URL : <https://arxiv.org/abs/2203.06631>.
- ROY, Swarnabha et al. (2022). « IoT Security and Computation Management on a Multi-Robot System for Rescue Operations Based on a Cloud Framework ». In : *Sensors* 22.15. ISSN : 1424-8220. DOI : [10.3390/s22155569](https://doi.org/10.3390/s22155569). URL : <https://www.mdpi.com/1424-8220/22/15/5569>.
- SAINATH, Tara N. et al. (2019). *Two-Pass End-to-End Speech Recognition*. arXiv : [1908.10992](https://arxiv.org/abs/1908.10992) [cs.CL]. URL : <https://arxiv.org/abs/1908.10992>.
- SALAM, Hanan et Mohamed CHETOUANI (2015). « Engagement detection based on mutli-party cues for human robot interaction ». In : *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, p. 341-347. DOI : [10.1109/ACII.2015.7344593](https://doi.org/10.1109/ACII.2015.7344593).
- SANDINI, Giulio et al. (2018). « Social Cognition for Human-Robot Symbiosis—Challenges and Building Blocks ». In : *Frontiers in Neurorobotics* 12. ISSN : 1662-5218. DOI : [10.3389/fnbot.2018.00034](https://doi.org/10.3389/fnbot.2018.00034). URL : <https://www.frontiersin.org/articles/10.3389/fnbot.2018.00034>.
- SATAKE, Satoru et al. (2009). « How to approach humans?-strategies for social robots to initiate interaction ». In : *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, p. 109-116. DOI : [10.1145/1514095.1514117](https://doi.org/10.1145/1514095.1514117).
- SCHEFLEN, Albert E (1964). *The Significance of Posture in Communication Systems*. University of Pennsylvania.
- SCHUSTER, M. et K.K. PALIWAL (1997). « Bidirectional recurrent neural networks ». In : *IEEE Transactions on Signal Processing* 45.11, p. 2673-2681. DOI : [10.1109/78.650093](https://doi.org/10.1109/78.650093).
- SERBAN, Iulian et al. (2017). « The Octopus Approach to the Alexa Competition : A Deep Ensemble-based Socialbot ». In : URL : <https://api.semanticscholar.org/CorpusID:108283712>.
- SHAMSHIRBAND, Shahab et al. (2021). « A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues ». In : *Journal of Biomedical Informatics* 113, p. 103627. ISSN : 1532-0464. DOI : <https://doi.org/10.1016/j.jbi.2020.103627>. URL : <https://www.sciencedirect.com/science/article/pii/S1532046420302550>.
- SHAYAN, Amir Mehdi et al. (2016). « RoboParrot 2.0: A multi-purpose social robot ». In : *2016 4th International Conference on Robotics and Mechatronics (ICROM)*, p. 422-427. DOI : [10.1109/ICRoM.2016.7886775](https://doi.org/10.1109/ICRoM.2016.7886775).

- SHIOMI, Masahiro, Takayuki KANDA et al. (2010). « A larger audience, please! — Encouraging people to listen to a guide robot ». In : *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, p. 31-38. DOI : [10.1109/HRI.2010.5453270](https://doi.org/10.1109/HRI.2010.5453270).
- SHIOMI, Masahiro, Daisuke SAKAMOTO et al. (sept. 2011). « Field Trial of a Networked Robot at a Train Station ». In : *I. J. Social Robotics* 3, p. 27-40. DOI : [10.1007/s12369-010-0077-4](https://doi.org/10.1007/s12369-010-0077-4).
- SHUM, Stephen H. et al. (2013). « Unsupervised Methods for Speaker Diarization: An Integrated and Iterative Approach ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 21.10, p. 2015-2028. DOI : [10.1109/TASL.2013.2264673](https://doi.org/10.1109/TASL.2013.2264673).
- SIDNER, Candace, Christopher LEE et Neal LESH (déc. 2003). « Engagement when looking: behaviors for robots when collaborating with people ». In.
- SIDNER, CANDACE L. et DZIKOVSKA, MYROSLAVA (2005). « A First Experiment in Engagement for Human-Robot Interaction in Hosting Activities ». In : *Advances in Natural Multimodal Dialogue Systems*. Sous la dir. de VAN KUPPEVELT, JAN C. J., DYBKJÆR, LAILA et BERNSEN, NIELS OLE. Dordrecht : Springer Netherlands, 55-76. ISBN : 978-1-4020-3933-1.
- SIGURDSSON, Gunnar A. et al. (2016). « Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding ». In : *ArXiv e-prints*. eprint : [1604.01753](https://arxiv.org/abs/1604.01753). URL : <http://arxiv.org/abs/1604.01753>.
- SIMONYAN, Karen et Andrew ZISSERMAN (2015). « Very deep convolutional networks for large-scale image recognition ». In : *International Conference on Learning Representations*.
- SIRITHUNGE, Chapa, Buddhika JAYASEKARA et Chandima D. PATHIRANA (juin 2019). « Proactive Robots With the Perception of Nonverbal Human Behavior: A Review Publisher: IEEE ». In : *IEEE Access* 7. DOI : [10.1109/ACCESS.2019.2921986](https://doi.org/10.1109/ACCESS.2019.2921986).
- SISBOT, Emrah, Raquel ROS et Rachid ALAMI (sept. 2011). « Situation assessment for human-robot interactive object manipulation ». In : p. 15-20. DOI : [10.1109/ROMAN.2011.6005258](https://doi.org/10.1109/ROMAN.2011.6005258).
- SKANTZE, Gabriel, Anna HJALMARSSON et Catharine OERTEL (nov. 2014). « Turn-taking, Feedback and Joint Attention in Situated Human-Robot Interaction ». In : *Speech Communication* 65. DOI : [10.1016/j.specom.2014.05.005](https://doi.org/10.1016/j.specom.2014.05.005).
- SOKOLOVA, Marina et Guy LAPALME (2009). « A systematic analysis of performance measures for classification tasks ». In : *Information Processing & Management* 45.4, p. 427-437. ISSN : 0306-4573. DOI : <https://doi.org/10.1016/j.ipm.2009.03.002>. URL : <https://www.sciencedirect.com/science/article/pii/S0306457309000259>.
- SPECIAN, Andrew et al. (2022). « Quori: A Community-Informed Design of a Socially Interactive Humanoid Robot ». In : *IEEE Transactions on Robotics* 38.3, p. 1755-1772. DOI : [10.1109/TR0.2021.3111718](https://doi.org/10.1109/TR0.2021.3111718).
- SZEGEDY, Christian et al. (2015). « Going deeper with convolutions ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 1-9.
- TADAS BALTRUŠAITIS, Marwa Mahmoud et Peter ROBINSON (2015). « Cross-dataset learning and person-specific normalisation for automatic Action Unit detection ». In : *IEEE International Conference on Automatic Face and Gesture Recognition*.

- TANAKA, F. et al. (2015). « Pepper learns together with children: Development of an educational application ». In : *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, p. 270-275. DOI : [10.1109/HUMANOIDS.2015.7363546](https://doi.org/10.1109/HUMANOIDS.2015.7363546).
- TARIMAN, Joseph et al. (sept. 2009). « Validation and testing of the Acceptability E-scale for Web-based patient-reported outcomes in cancer care ». In : *Applied nursing research : ANR* 24, p. 53-8. DOI : [10.1016/j.apnr.2009.04.003](https://doi.org/10.1016/j.apnr.2009.04.003).
- Tessa Robot - Zorgrobot voor zelfregie en zelfstandig wonen* (s. d.). Accessed: 2023-12-04. Tinybots. URL : <https://www.tinybots.nl/>.
- TOISOUL, Antoine et al. (jan. 2021). « Estimation of continuous valence and arousal levels from faces in naturalistic conditions ». In : *Nature Machine Intelligence* 3. DOI : [10.1038/s42256-020-00280-0](https://doi.org/10.1038/s42256-020-00280-0).
- TOMASELLO, Michael et al. (2005). « Understanding and sharing intentions: The origins of cultural cognition ». In : *Behavioral and Brain Sciences* 28.5, p. 675-691. DOI : [10.1017/S0140525X05000129](https://doi.org/10.1017/S0140525X05000129).
- TOOLS, Enchanted (2023). *Enchanted Tools: Changing the Face of Robotics*. URL : <https://enchanted.tools/>.
- TRANter, S.E. et D.A. REYNOLDS (2006). « An overview of automatic speaker diarization systems ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 14.5, p. 1557-1565. DOI : [10.1109/TASL.2006.878256](https://doi.org/10.1109/TASL.2006.878256).
- URQUIZA-HAAS, Esmeralda G. et Kurt KOTRSCHAL (2015). « The mind behind anthropomorphic thinking: attribution of mental states to other species ». In : *Animal Behaviour* 109, p. 167-176. ISSN : 0003-3472. DOI : <https://doi.org/10.1016/j.anbehav.2015.08.011>. URL : <https://www.sciencedirect.com/science/article/pii/S0003347215003085>.
- VAPNIK, Vladimir (1963). « Pattern recognition using generalized portrait method ». In : *Automation and Remote Control* 24, p. 774-780.
- VASWANI, Ashish et al. (2017). « Attention is All You Need ». In : *Advances in Neural Information Processing Systems* 30.
- VERSCHURE, P.F., C.M. PENNARTZ et G. PEZZULO (2014). « The why, what, where, when and how of goal-directed choice: neuronal and computational principles ». In : *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1655, p. 20130483. DOI : [10.1098/rstb.2013.0483](https://doi.org/10.1098/rstb.2013.0483).
- VERSCHURE, Paul, Thomas VOEGTLIN et Rodney DOUGLAS (nov. 2003). « Environmentally mediated synergy between perception and behavior in mobile robots ». In : *Nature* 425, p. 620-4. DOI : [10.1038/nature02024](https://doi.org/10.1038/nature02024).
- VILLALBA, Jesús et al. (sept. 2019). « State-of-the-Art Speaker Recognition for Telephone and Video Speech: The JHU-MIT Submission for NIST SRE18 ». In : *Interspeech*, p. 1488-1492. DOI : [10.21437/Interspeech.2019-2713](https://doi.org/10.21437/Interspeech.2019-2713).
- VINCIARELLI, Alessandro, Maja PANTIC et al. (oct. 2008). « Social Signal Processing: State-of-the-Art and Future Perspectives of an Emerging Domain ». In : *MM'08 - Proceedings of the 2008 ACM International Conference on Multimedia, with co-located Symposium and Workshops*. DOI : [10.1145/1459359.1459573](https://doi.org/10.1145/1459359.1459573).
- VINCIARELLI, Alessandro, H. SALAMIN et M. PANTIC (juin 2009). « Social Signal Processing: Understanding social interactions through nonverbal behavior analysis ». In : p. 42-49. DOI : [10.1109/CVPR.2009.5204290](https://doi.org/10.1109/CVPR.2009.5204290).

- VIOLA, P. et M. JONES (2001). « Rapid object detection using a boosted cascade of simple features ». In : *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. T. 1, p. I-I. DOI : [10.1109/CVPR.2001.990517](https://doi.org/10.1109/CVPR.2001.990517).
- WANG, Chien-Yao, Alexey BOCHKOVSKIY et Hong-Yuan Mark LIAO (2021). *Scaled-YOLOv4: Scaling Cross Stage Partial Network*. arXiv : [2011.08036](https://arxiv.org/abs/2011.08036) [cs.CV].
- (2022). *YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors*. arXiv : [2207.02696](https://arxiv.org/abs/2207.02696) [cs.CV].
- WANG, Chien-Yao, I-Hau YEH et Hong-Yuan Mark LIAO (2021). *You Only Learn One Representation: Unified Network for Multiple Tasks*. arXiv : [2105.04206](https://arxiv.org/abs/2105.04206) [cs.CV].
- WANG, Jinjiang et al. (2018). « Deep learning for smart manufacturing: Methods and applications ». In : *Journal of Manufacturing Systems* 48. Special Issue on Smart Manufacturing, p. 144-156. ISSN : 0278-6125. DOI : <https://doi.org/10.1016/j.jmsy.2018.01.003>. URL : <https://www.sciencedirect.com/science/article/pii/S0278612518300037>.
- WANG, Quan et al. (2017). « Speaker Diarization with LSTM ». In : *arXiv.org*. DOI : [10.48550/ARXIV.1710.10468](https://doi.org/10.48550/ARXIV.1710.10468). URL : <https://arxiv.org/abs/1710.10468>.
- WANG, Yiming et Andrea CAVALLARO (2016). « Prioritized target tracking with active collaborative cameras ». In : *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, p. 131-137. DOI : [10.1109/AVSS.2016.7738066](https://doi.org/10.1109/AVSS.2016.7738066).
- WEBB, Nicola, Manuel GIULIANI et Séverin LEMAIGNAN (2022). « Measuring Visual Social Engagement from Proxemics and Gaze ». In : *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, p. 757-762. DOI : [10.1109/RO-MAN53752.2022.9900801](https://doi.org/10.1109/RO-MAN53752.2022.9900801).
- WOJKE, Nicolai, Alex BEWLEY et Dietrich PAULUS (2017). *Simple Online and Realtime Tracking with a Deep Association Metric*. arXiv : [1703.07402](https://arxiv.org/abs/1703.07402) [cs.CV].
- WYER, Robert S. et Donal E. CARLSTON (1979). *Social cognition, Inference, and attribution*. HALSTED PRESS.
- XIONG, Zichao et al. (2021). « Multi-Speaker Tracking by Fusing Audio and Video Information ». In : *2021 IEEE Statistical Signal Processing Workshop (SSP)*, p. 321-325. DOI : [10.1109/SSP49050.2021.9513767](https://doi.org/10.1109/SSP49050.2021.9513767).
- YE, Qinghao et al. (2024). *mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality*. arXiv : [2304.14178](https://arxiv.org/abs/2304.14178) [cs.CL].
- YORKE-SMITH, NEIL et al. (2012). « THE DESIGN OF A PROACTIVE PERSONAL AGENT FOR TASK MANAGEMENT ». In : *International Journal on Artificial Intelligence Tools* 21.01, p. 1250004. DOI : [10.1142/s0218213012500042](https://doi.org/10.1142/s0218213012500042). URL : <https://doi.org/10.1142/s0218213012500042>.
- YOUSUF, Mohammad A. et al. (2013). « How to move towards visitors: A model for museum guide robots to initiate conversation ». In : *2013 IEEE RO-MAN*, p. 587-592. DOI : [10.1109/ROMAN.2013.6628543](https://doi.org/10.1109/ROMAN.2013.6628543).
- ZHANG, Kaipeng et al. (2016). « Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks ». In : *CoRR* abs/1604.02878. arXiv : [1604.02878](https://arxiv.org/abs/1604.02878). URL : <http://arxiv.org/abs/1604.02878>.
- ZHAO, Wayne Xin et al. (2023). *A Survey of Large Language Models*. arXiv : [2303.18223](https://arxiv.org/abs/2303.18223) [cs.CL]. URL : <https://arxiv.org/abs/2303.18223>.

ZHOU, Xingyi, Dequan WANG et Philipp KRÄHENBÜHL (2019). *Objects as Points*. arXiv : [1904.07850](https://arxiv.org/abs/1904.07850) [cs.CV].