Détection de marqueurs affectifs et attentionnels de personnes âgées en interaction avec un robot

Fan Yang

NNT : 2015SACLS081

THESE DE DOCTORAT
DE
L'UNIVERSITE PARIS-SACLAY
PREPAREE A
L'UNIVERSITE PARIS-SUD

ECOLE DOCTORALE N° 580
Sciences et technologies de l'information et de la communication

Spécialité de doctorat : Informatique

Par

M. Fan YANG

Détection de marqueurs affectifs et attentionnels de personnes âgées en interaction avec un robot

Thèse présentée et soutenue à Orsay (France), le 23/10/2015 :

Composition du Jury :

Mme VILNAT Anne	Professeur (Paris Sud,LIMSI-CNRS)	Présidente & Examinatrice
M. SCHULLER Björn W.	Professeur (Imperial College London)	Rapporteur
M. CHETOUANI Mohamed	Professeur (UPMC, ISIR)	Rapporteur
M. QUENOT Georges	Directeur de recherche (LIG-CNRS)	Examinateur
M. BARRAS Claude	Maître de conférences (Paris Sud,LIMSI-CNRS)	Directeur de thèse
Mme DEVILLERS Laurence	Professeur (Paris Sorbonne IV, LIMSI-CNRS)	Co-directeur de thèse

Remerciements

Je voudrais d'abord remercier mes directeurs de thèse Claude BARRAS et Laurence DEVILLERS de m'avoir offert l'opportunité de faire cette thèse, ainsi que pour leurs suggestions et encouragements constants.

Je remercie les membres de mon jury, et en particulier mes rapporteurs de thèse Mohamed CHETOUANI et Björn W. SCHULLER d'avoir accordé du temps à mes travaux.

Je remercie mon collègue et mon ami Mohamed SEHILI avec qui j'ai fait beaucoup de collaboration de recherche.

Je remercie aussi Agnès, Clément C., Clément G. Guillaume, Julietta, Lucile, Mariette, et Marie, les autres membres de l'équipe « Dimensions affectives et sociales dans les interactions orales » pour leur collaboration, leur soutien et leur amitié.

Je souhaite aussi remercier les 27 personnes âgées qui ont participé à la collection du corpus ROMEO2, sans qui cette étude n'aurait pas pu être réalisée.

Finalement, je tiens à remercier tout particulièrement ma famille, pour m'avoir permis de faire mes études à l'étranger pour leur soutien à la fois affectif et financier.

Mes remerciements s'adressent également à tous les « limsiens » pour leur accueil chaleureux.

Résumé

Ces travaux de thèse portent sur la détection audio-visuelle de marqueurs affectifs (rire et sourire) et attentionnels de personnes âgées en interaction sociale avec un robot.

Pour comprendre efficacement et modéliser le comportement des personnes très âgées en présence d'un robot, des données pertinentes sont nécessaires. J'ai participé à la collection d'un corpus de personnes âgées notamment pour l'enregistrement des données visuelles. Le système utilisé pour contrôler le robot est un magicien d'Oz, plusieurs scénarios de conversation au quotidien ont été utilisés pour encourager les gens à coopérer avec le robot. Ces scénarios ont été élaborés dans le cadre du projet ROMEO2[1] avec l'association Approche[2].

Nous avons décrit tout d'abord le corpus recueilli qui contient 27 sujets de 85 ans en moyenne pour une durée totale de 9 heures, les annotations et nous avons discuté des résultats obtenus à partir de l'analyse des annotations et de deux questionnaires.

Ma recherche se focalise ensuite sur la détection de l'attention et la détection de rire et de sourire. Les motivations pour la détection de l'attention consistent à détecter quand le sujet ne s'adresse pas au robot et à adapter le comportement du robot à la situation. Après avoir considéré les difficultés liées aux personnes âgées et les résultats d'analyse obtenus par l'étude des annotations du corpus, nous nous intéressons à la rotation de la tête au niveau de l'indice visuel et à l'énergie et la qualité de voix pour la détection du destinataire de la parole. La détection de rire et sourire peut être utilisée pour l'étude sur le profil du locuteur et de ses émotions. Mes intérêts se concentrent sur la détection de rire et sourire dans la modalité visuelle et la fusion des informations audio-visuelles afin d'améliorer la performance du système automatique.

Les expressions sont différentes des expressions actées ou posés à la fois en apparence et en temps de réaction. La conception d'un système qui marche sur les données réalistes des personnes âgées est encore plus difficile à cause de plusieurs difficultés à envisager telles que le manque de données pour l'entrainement du modèle statistique, l'influence de la texture faciale et de la façon de sourire pour la détection visuelle, l'influence de la qualité vocale pour la détection auditive, la variété du temps de réaction, le niveau de compréhension auditive, la perte de la vue des personnes âgées, etc.

[1] http://projetromeo.com

[2] http://www.approche-asso.com/

Les systèmes de détection de la rotation de la tête, de la détection de l'attention et de la détection de rire et sourire sont évalués sur le corpus ROMEO2 et partiellement évalués (détections visuelles) sur les corpus standard Pointing04 et GENKI-4K pour comparer avec les scores des méthodes de l'état de l'art.

Nous avons également trouvé une corrélation négative entre la performance de détection de rire et sourire et le nombre d'évènement de rire et sourire pour le système visuel et le système audio-visuel. Ce phénomène peut être expliqué par le fait que les personnes âgées qui sont plus intéressées par l'expérimentation rient plus souvent et sont plus à l'aise donc avec des poses variées. La variété des poses et le manque de données correspondantes amènent des difficultés pour la reconnaissance de rire et de sourire pour les systèmes statistiques.

Les expérimentations montrent que la rotation de la tête peut être efficacement utilisée pour détecter la perte de l'attention du sujet dans l'interaction avec le robot. Au niveau de la détection de l'attention, le potentiel d'une méthode en cascade qui utilise les modalités d'une manière complémentaire est montré. Cette méthode donne de meilleurs résultats que le système auditif seul. Pour la détection de rire et sourire, en suivant le même protocole « Leave-one-out », la fusion des deux systèmes monomodaux améliore aussi significativement la performance par rapport à un système monomodal au niveau de l'évaluation segmentale.

Mots clé : attention, marqueur affectif, rire et sourire, détection multimodale, personne âgée, corpus réaliste, interaction sociale avec robot

Abstract

This thesis work focuses on audio-visual detection of emotional (laugh and smile) and attentional markers for elderly people in social interaction with a robot.

To effectively understand and model the pattern of behavior of very old people in the presence of a robot, relevant data are needed. I participated in the collection of a corpus of elderly people in particular for recording visual data. The system used to control the robot is a Wizard of Oz, several daily conversation scenarios were used to encourage people to interact with the robot. These scenarios were developed as part of the ROMEO2[3] project with the Approche[4] association.

We described at first the corpus collected which contains 27 subjects of 85 years' old on average for a total of 9 hours, annotations and we discussed the results obtained from the analysis of annotations and two questionnaires.

My research then focuses on the attention detection and the laughter and smile detection. The motivations for the attention detection are to detect when the subject is not addressing to the robot and adjust the robot's behavior to the situation. After considering the difficulties related to the elderly people and the analytical results obtained by the study of the corpus annotations, we focus on the rotation of the head at the visual index and energy and quality vote for the detection of the speech recipient. The laughter and smile detection can be used to study on the profile of the speaker and her emotions. My interests focus on laughter and smile detection in the visual modality and the fusion of audio-visual information to improve the performance of the automatic system.

Spontaneous expressions are different from posed or acted expression in both appearance and timing. Designing a system that works on realistic data of the elderly is even more difficult because of several difficulties to consider such as the lack data for training the statistical model, the influence of the facial texture and the smiling pattern for visual detection, the influence of voice quality for auditory detection, the variety of reaction time, the level of listening comprehension, loss of sight for elderly people, etc.

The systems of head-turning detection, attention detection and laughter and smile detection are evaluated on ROMEO2 corpus and partially evaluated (visual detections)

[3] http://projetromeo.com

[4] http://www.approche-asso.com/

on standard corpus Pointing04 and GENKI-4K to compare with the scores of the methods on the state of the art.

We also found a negative correlation between laughter and smile detection performance and the number of laughter and smile events for the visual detection system and the audio-visual system. This phenomenon can be explained by the fact that elderly people who are more interested in experimentation laugh more often and therefore perform more various poses. The variety of poses and the lack of corresponding data bring difficulties for the laughter and smile recognition for our statistical systems.

The experiments show that the head-turning can be effectively used to detect the loss of the subject's attention in the interaction with the robot. For the attention detection, the potential of a cascade method using both methods in a complementary manner is shown. This method gives better results than the audio system. For the laughter and smile detection, under the same leave-one-out protocol, the fusion of the two monomodal systems significantly improves the performance of the system at the segmental evaluation.

Keywords: attention, affect burst, laughter and smile, multimodal detection, elderly people, realistic corpus, social interaction with robot

Table des matières

Liste des Tableaux

Liste des figures

Glossaire

- AGGIR : Autonomie Gérontologie Groupes Iso-Ressources
- Affect Bursts : Expressions émotionnelles spontanées de courte durée
- BER : Balanced Error Rate (ratio d'erreur balancé)
- EHPAD : Etablissement d'Hébergement pour Personnes Agées Dépendantes
- FACS : Facial Action Coding System
- GMM : Gaussian Mixture Models
- GIR : Groupe Iso-Ressources allant de la dépendance la plus lourde (GIR 1) à l'absence de perte d'autonomie (GIR 6)
- HD : Haute Définition
- HMM : Hidden Markov Models
- LBP : Local binary patterns
- LBP-TOP : Local Binary Pattern on Three Orthogonal Planes
- OCEAN : Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism
- PLP : Perceptual Linear Prediction (coding features)
- RGB : Red(rouge), Green(vert), Blue(bleu)
- SVM : Support Vector Machines
- WoZ : Wizard of Oz (Magicien d'Oz)

Introduction

Le sujet de ma thèse est la « Détection de marqueurs affectifs et attentionnels de personnes âgées en interaction avec un robot », qui fait partie du domaine de l'affective computing, décrit dans les travaux de R. Picard en 1997 [102]. Les recherches dans le domaine de l'"affective computing" se focalisent sur la modélisation informatique des émotions et plus largement des comportements affectifs. De plus en plus de recherche portent sur les marqueurs sociaux et affectifs. Dans mon étude de thèse, je me suis concentré sur la détection visuelle de l'attention, de rire et de sourire de personnes âgées, et la fusion audio-visuelle pour la détection.

Nos intérêts des études couvrent alors :

- L'extraction des marqueurs affectifs et sociaux dans les interactions humain-robot et plus particulièrement:
  - la détection de l'attention
  - les marqueurs affectifs tels que le rire et le sourire
- La Combinaison d'indices ou de modalités audio et vidéo : fusion précoce ou tardive
- Les données "real-life": Corpus ROMEO2 d'interaction avec les personnes âgées (NAO)

Corpus ROMEO2

La plupart des corpus émotionnels existants sont actés [65, 79, 97, 139], mais peu de corpus réalistes existent. Il y en a encore moins quand il s'agit de données réalistes avec des personnes âgées. Dans le but d'avoir un corpus de données réalistes avec des marqueurs affectifs et sociaux des personnes âgées, j'ai participé à la collection d'un corpus de personnes âgées notamment pour l'enregistrement des données visuelles. Le système utilisé pour contrôler le robot est un magicien d'Oz, plusieurs scénarios de conversation au quotidien ont été utilisés pour encourager les gens à coopérer avec le robot. Ces scénarios ont été élaborés dans le cadre du projet ROMEO2[5] avec l'association Approche[6]. La mission de l'association Approche est de promouvoir les nouvelles technologies au service des personnes en situations de handicap. Le but du projet ROMEO et ROMEO2 est de développer un robot humanoïde qui peut agir comme un assistant d'accompagnement pour les personnes souffrant de perte d'autonomie. Dans cette perspective, le robot est en mesure d'aider une personne dans

[5] http://projetromeo.com

[6] http://www.approche-asso.com/

ses tâches quotidiennes quand elle est seule. Le but de cette étude est de concevoir un système interactif affectif entraîné avec des marqueurs interactionnels, émotionnels et de la personnalité. J'ai participé à ce travail au sein de l'équipe « Dimensions affectives et sociales dans les interactions orales » sur la partie détection visuelle du rire, du sourire et de l'attention ainsi que sur la fusion audio-visuelle tardive en collaboration avec Mohamed SEHILI.

L'ensemble du corpus ROMEO2 collecté contient 27 sujets de 85 ans en moyenne pour une durée totale de 9 heures. Afin d'étudier précisément le comportement des personnes âgées, j'ai conçu le schéma d'annotation audio et vidéo avec les chercheurs de l'équipe, annoté une partie du corpus, relu et corrigé les annotations des autres annotateurs, complété les résultats d'analyse des corrélations entre les questionnaires de satisfaction et de personnalité, et participé aux analyses sur les corrélations entre questionnaires et annotations. Plusieurs phénomènes liés au comportement des personnes âgées en interaction avec le robot ont été trouvés, par exemple, les personnes âgées se sont souvent adressées à l'expérimentateur qui était assis à leur droite. Cette constatation nous a amené à étudier des signes de perte de l'attention (e.g. tourner la tête) des personnes dans l'interaction.

Défis des systèmes automatiques

L'objectif de ma thèse est l'étude de l'interaction affective et sociale entre des personnes âgées et un robot à partir du corpus ROMEO2, le choix des indices de comportement à étudier dans le système automatique doit considérer plusieurs facteurs tels que l'importance des indices pour le déroulement de l'expérimentation, l'apprentissage de l'état mental ou émotionnel du sujet, l'influence sur la stratégie de communication du robot, la quantité d'évènements annotés pour l'entrainement des modèles statistiques, etc. Ma recherche se focalise sur la détection de l'attention et la détection de rire et de sourire.

La plupart des chercheurs ont testé et validé leurs méthodes de détection sur les corpus actés ou posés [38, 59, 73, 120]. M. Valstar et de ses collègues [142] affirment que les expressions spontanées sont différentes des expressions actées ou posés à la fois en apparence et en temps de réaction. Cela signifie que les méthodes utilisées pour la reconnaissance des expressions actées ou posées pourraient ne pas fonctionner correctement sur les expressions réalistes. De plus, la conception d'un système qui marche sur les données réalistes des personnes âgées est encore plus difficile à cause du manque de données pour l'entrainement du modèle statistique, de l'influence de la texture faciale et de la façon de sourire pour la détection visuelle, de l'influence de la qualité vocale pour la détection auditive, de la variété du temps de réaction, du niveau de compréhension auditive, de la perte de la vue des personnes âgées, etc. Tous ces défis liés à la reconnaissance automatique sur le corpus réaliste des personnes âgées en interaction sociale avec robot sont envisagés. Ces difficultés également exigent,

hors le corpus ROMEO2, une évaluation du système sur un corpus standard pour comparer avec la performance des méthodes de l'état de l'art.

Détection de l'attention

Dans de nombreuses interactions sociales humain-robot, le sujet est très susceptible d'interagir avec d'autres êtres humains présents dans la même pièce et perd temporellement le focus sur l'interaction principale avec le robot. Cette interaction humain-humain peut être une très brève interaction ou une assez longue discussion. Les motivations de la détection de l'attention consistent à percevoir quand le sujet ne s'adresse pas au robot et à adapter le comportement du robot à la situation. Dans de nombreux ouvrages, le suivi de regard et la technique de localisation audio sont utilisés pour détecter l'attention du sujet. Après avoir considéré les difficultés liées aux personnes âgées et les résultats d'analyse obtenus par l'étude des annotations du corpus, nous nous intéressons à la rotation de la tête au niveau de l'indice visuel, l'énergie et la qualité de voix pour la détection du destinataire de la parole. Un sous-ensemble du corpus ROMEO 2 et le corpus standard Pointing04 [53] (pour détection visuelle) est utilisé pour l'expérimentation du système automatique.

Détection de rire et sourire

Les marqueurs affectifs [109] jouent un rôle important dans l'interaction sociale non verbale, parmi lesquels le rire et le sourire sont parmi les marqueurs sociaux les plus importants de l'interaction homme-robot sociale. Ils ne contiennent pas seulement des informations affectives mais ils peuvent également révéler la stratégie de communication de locuteur. Dans le contexte de l'interaction homme robot, un système de détection automatique de rire et sourire peut donc aider le robot à adapter son comportement au profil de l'utilisateur donné en adoptant une stratégie de communication plus pertinente. Même si de nombreuses études intéressantes sur la détection de rire et sourire ont été menées, peu d'entre elles portaient sur les personnes âgées. Les données de personnes âgées sont relativement rares et portent souvent un défi important pour le système automatique de détection du rire et sourire en raison de l'influence de rides faciales pour la reconnaissance visuelle et la faible qualité de voix pour la reconnaissance auditive.

Mes intérêts se concentrent sur la détection de rire et sourire dans la modalité visuelle et la fusion des informations des modalités audio et visuelles afin d'améliorer la performance du système automatique. La forte corrélation entre la relaxation au niveau de l'expérience pendant l'expérimentation des sujets et leur nombre de sourires et rire annotées lors d'une interaction est montrée, les corrélations entre le nombre d'évènements et la performance du système automatique sont également trouvées.

Apports de l'étude

Cette étude a contribué à plusieurs aspects de d'étude affective et sociale des personnes âgées :

- Participation à la collecte notamment pour l'enregistrement de vidéo du corpus ROMEO2
- Conception du schéma d'annotation avec les collègues et participation à l'annotation, la vérification et la correction des annotations
- Analyse des corrélations entre l'âge, la personnalité, l'autonomie et les comportements des personnes âgées à partir des questionnaires et des annotations du corpus ROMEO2
- Analyse de la dimension sociale de la rotation de la tête
- Conception et réalisation d'une méthode visuelle non-statistique pour la détection de rotation de la tête ; l'évaluation sur le corpus standard Pointing04 et l'évaluation au niveau de frame et segment sur une partie du corpus ROMEO2
- Fusion de la détection visuelle de la rotation de la tête et la détection auditive de la qualité de voix (conception et réalisation du système audio par mon collègue Mohamed Sehili) pour la détection de l'attention ; l'évaluation segmentale sur une partie du corpus ROMEO2
- Détection de rire et sourire dans la modalité visuelle
- Fusion des informations des modalités audio-visuelles afin d'améliorer la performance du système automatique (conception et réalisation du système audio par mon collègue Mohamed Sehili) ; l'évaluation au niveau des frames, du segment (séquence de frames) et « in the wild » (sans segmentation a priori) sur une partie du corpus ROMEO2
- Analyse des corrélations statistiques entre la performance des systèmes de détection de rire et sourire pour les personnes âgées et les questionnaires, les annotations

Organisation du Manuscrit

Ce manuscrit s'organise en trois parties :

La première partie présente une étude de l'état de l'art sur le domaine de l'« Affective Computing », particulièrement pour la détection des marqueurs affectifs et attentionnels. Cette partie est divisée en deux chapitres. Le chapitre 1 s'intéresse à la présentation de la notion de l'« Affective Computing », les travaux dans la recherche de la détection des émotions, de l'attention, de rire et sourire et les méthodes appliquées sont présentés. De plus, un résumé des corpus existants est également fourni. Le chapitre 2 se focalise sur la présentation des méthodes de traitement

d'image, du modèle de classification statistique, de l'approche en fusion multimodale et des mesures d'évaluation de la performance du système utilisé dans mes expérimentations.

La deuxième partie présente nos travaux de collecte, d'annotation et d'analyse du corpus ROMEO2. Le protocole de collecte de données et les scénarios d'interaction conçus ainsi que le corpus recueilli seront présentés dans le chapitre 3. Nous décrirons ensuite le schéma d'annotation et la quantité d'évènements annotés dans le paragraphe 3.9. Nous discuterons des résultats obtenus à partir de l'analyse des annotations et de deux questionnaires (un questionnaire de satisfaction, et le Big-five qui est un questionnaire sur la personnalité) dans le chapitre 4.

La dernière partie présente les systèmes automatiques pour la détection des marqueurs affectifs et attentionnels de personnes âgées en interaction avec un robot. Le chapitre 5 présente une méthode conçue à la base de plusieurs détecteurs de visage de différente orientation pour la détection de la rotation de la tête. Le système est évalué sur une partie du corpus ROMEO2 et le corpus standard Pointing04. De plus, la dimension sociale de la rotation de la tête est également étudiée dans ce chapitre. Le chapitre 6 utilise la méthode de détection de rotation de la tête présentée dans le chapitre précédant en combinant avec une méthode de modalité auditive pour la détection de l'attention. Le système est testé sur une partie du corpus ROMEO2. Le chapitre 7 présente un système audio-visuel pour la détection de rire et sourire. L'ensemble du système est évalué sur une partie du corpus ROMEO2, la détection visuelle est également évaluée sur le corpus standard GENKI-4K. De plus, les corrélations statistiques entre la performance des systèmes et les questionnaires et les annotations sont analysées.

La conclusion est une synthèse du travail réalisé dans la thèse, et ouvre des perspectives sur les défis encore à réaliser pour les applications de détection de marqueurs affectifs et attentionnels de personnes âgées en interaction avec un robot.

Partie I : Etat de l'art

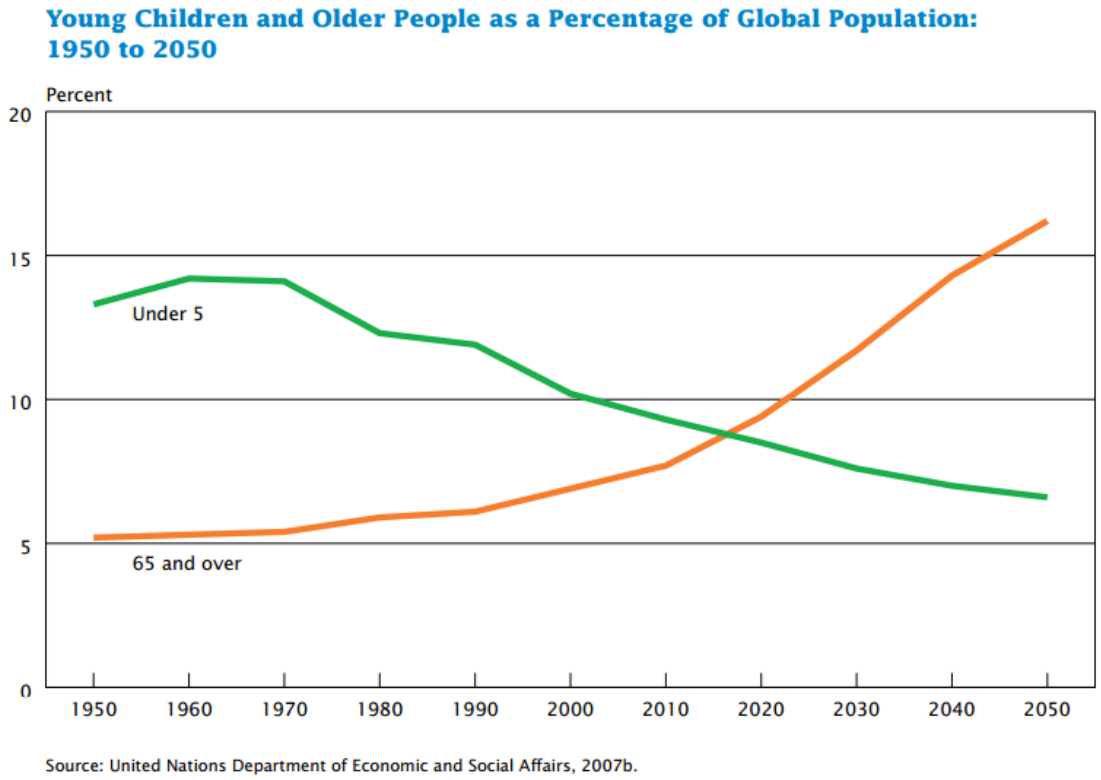

Figure 1 : Distribution des pourcentages des populations d'enfants et de personnes âgées au niveau mondial. (Source de l'image: « An Aging World: 2008 », International Population Reports, United States Census Bureau, 2008)

Avec le développement de la technologie et la prolongation de la durée de vie, le phénomène du vieillissement démographique devient de plus en plus significatif. Selon l'estimation du gouvernement américain en 2007 illustrée dans la Figure 1, la population des personnes âgées au-dessus de 65 ans va dépasser 16% de la population totale au niveau mondial en 2050. En France, la situation est encore plus grave. Selon les données statistiques fournies par l'Institut national de la statistique et des études économiques (Insee[7]) montrées dans le Tableau 1, au début de l'année 2015, il y a déjà plus de 18% des français au-dessus de 65 ans. Selon l'estimation du « United States Census Bureau [8]», la population du groupe d'âge supérieur à 65 ans pour l'Europe de l'Ouest en 2040 augmentera à 28,1% dont 15,0% pour les gens de plus de 75 ans et 9,3% pour les gens de plus de 80 ans. Les personnes âgées, en tant que groupe très important de personnes en perte d'autonomie, ont attiré de plus en plus

[7] www.insee.fr

[8] http://www.census.gov/

l'attention de la communauté de recherche. Il y a également de plus en plus de projets européens et mondiaux qui se focalisent sur l'assistance aux personnes âgées, comme les projets ROMEO et ROMEO2 qui ont pour objectif de développer un robot humanoïde qui peut agir comme un assistant d'accompagnement pour les personnes souffrant de perte d'autonomie.

Tableau 1 : Distribution de la population selon l'âge en France au début de l'année 2015.

Age	Millions	%	% femmes
65+	12,19	18,4	57,5
20-64	37,76	56,9	50,8
<20	16,37	24,7	48,9
Total	66,32	100	51,6

Dans cette première partie du rapport de la thèse, un aperçu des travaux existants sur le domaine récent de l'« Affective Computing » qui se focalise sur l'étude affective des humains est présenté. La recherche dans le domaine « Affective Computing » n'est pas limitée à l'étude des émotions. A cause de la complexité des émotions humaines dans la vie réelle, de plus en plus de recherches portent sur l'étude des marqueurs affectifs et attentifs des humains. Dans cette partie de ma thèse, la présentation de l'état de l'art se focalise particulièrement sur la détection de l'attention, de rire et sourire ainsi que sur les outils de reconnaissance automatique correspondants.

Le chapitre 1 « Affective computing » présente d'abord les travaux liés à la détection des émotions, puis un ensemble de descriptions précises sur les méthodes audio et visuelles utilisées dans la communauté pour la détection de l'attention, du rire et du sourire ainsi que des corpus existants seront fournies en les adaptant à notre objectif de l'étude sur les personnes âgées.

Le chapitre 2 « Techniques d'analyse de l'image » se focalise sur la présentation des méthodes de traitement d'image, du modèle de classification statistique, de l'approche en fusion multimodale et des mesures d'évaluation de la performance du système utilisées dans mes expérimentations.

1 « Affective computing »

L'essentiel du domaine de l' « Affective Computing » (en français : informatique affective) établi par les travaux de Rosalind Picard [102] est de l'étude et la création de systèmes capables d'analyser, détecter, exprimer et modéliser les états affectifs dont les émotions, les marqueurs affectifs (affect bursts), etc.

1.1 Détection des émotions

Il existe 2 théories principales sur la représentation des émotions : la représentation par étiquettes verbales [46] et la représentation en dimensions abstraites [104]. L'idée de la première théorie est qu'il existe un nombre limité d'émotions primaires (joie, colère, tristesse, peur, dégoût, surprise), et que ces émotions primaires peuvent se combiner pour obtenir d'autres émotions. Par exemple l'amour est la somme des émotions de joie et d'acceptation. La deuxième théorie décrit l'état émotionnel selon plusieurs axes dimensionnels notamment : la valence (de positive à négative), l'activation (d'active à passive) et le pouvoir (potentiel de critique, de fort à faible). Par exemple la joie est considérée comme une émotion active et positive, la colère est considérée comme une émotion négative et active. En dehors des deux représentations des émotions principales, il existe également des théories hybrides [94, 103].

Selon Picard [102], la représentation des émotions comme des étiquettes verbales ou les dimensions abstraites est juste une question de choix qui ressemble beaucoup à la représentation de la lumière, laquelle peut être décrite en utilisant les théories des ondes et des particules. Le choix de la théorie dépend de l'explication recherchée. Une couleur peut être décrite par ses valeurs RGB ou par son nom, le choix dépend de l'application. Et comme la représentation des couleurs, dans l'article [105], les auteurs présentent une table de mots décrivant des émotions avec leurs coordonnées correspondantes dans les dimensions abstraites.

Les systèmes automatiques extraient les informations émotionnelles à partir des indices audio-visuels, particulièrement la voix et l'image faciale du locuteur mais aussi des gestes et des postures. Toutes ces informations sont utilisées pour entrainer des classifieurs statistiques. L'émotion peut être détectée selon la voix, le visage ou encore les gestes. Les premiers travaux en « affective computing » ont été menés au MIT par R. Picard en 1997 [102]. La recherche sur la détection des émotions est désormais très active : ce thème est maintenant largement étudié dans de nombreux laboratoires internationaux sur différents aspects, avec la reconnaissance d'émotions dans les flux vidéo, physiologiques, etc. L'utilisation des émotions dans les interactions homme-machine est un sujet d'actualité, objet de nombreuses recherches [13, 17, 37], dont la détection des émotions dans la voix [13, 28, 39, 111, 127], la reconnaissance d'émotions dans les flux vidéo [2, 16, 49, 87, 150], et multimodale [1,

113]. La prise en compte d'informations multimodales permet d'améliorer les scores de détection : par exemple l'information labiale améliore l'intelligibilité de la voix pour l'humain [81], cela peut aussi améliorer la performance de la détection de la voix par la machine [1]. Pour indication, les résultats au niveau état de l'art pour la catégorisation discrète des émotions à partir d'indices oraux sont d'environ 80% de bonne détection pour 2 émotions, 60% pour 4 [40].

Mais dans la réalité, les émotions humaines sont très complexes, même les humains ont des difficultés à distinguer et à classifier les émotions réalistes. La mesure d'inter-corrélation des annotations émotionnelles par plusieurs annotateurs n'est pas toujours fiable. De plus, la performance du système est facilement influencée par les conditions environnementales, par exemple : le bruit, l'écho, la texture du visage. Tous ces défis augmentent la difficulté de la détection des émotions dans des conditions « real-life », particulièrement pour les personnes âgées. Une détection des marqueurs affectifs « Affect bursts » sera une méthode complémentaire pour la reconnaissance affective.

1.2 « Affect bursts » – détection de rire et sourire

Le concept d'« Affect burst » porte sur les expressions émotionnelles spontanées de courte durée [143]. Selon Scherer, les « Affect bursts » sont définis dans son article [108] comme "Very brief, discrete, nonverbal expressions of affect in both face and voice as triggered by clearly identifiable events". Les « Affect bursts » jouent un rôle important dans l'interaction sociale non verbale, parmi lesquels le rire et le sourire sont les marqueurs sociaux les plus importants de l'interaction homme-robot sociale. Ils véhiculent beaucoup d'informations au cours de l'interaction humain-humain et humain-machine, telles que l'état émotionnel, la stratégie sociale de la communication et de la personnalité. Leur détection a également une grande variété d'applications telles que la mesure de satisfaction du client par rapport au produit, au jeu ou à l'interface d'utilisation, etc.

De nombreux travaux récents portent sur la recherche de sourire et de rire dans le domaine de l'informatique, en particulier dans l'interaction homme-machine. Des workshops consacrés au thème sont en cours organisées telles que l' « Interdisciplinary Workshop on Laughter and other Non-Verbal Vocalizations in Speech »[9] et des projets internationaux comme le projet « ILHAIRE»[10] [83, 90] ont également été lancés.

[9] https://laughterworkshop2015.wordpress.com/

[10] http://www.ilhaire.eu/

Le rire et le sourire,ont par nature beaucoup de points communs. Ils consistent en expressions du visage composées notamment par le mouvement de la bouche (et aussi des yeux), sont des expressions universelles au niveau culturel, de plus, ils sont souvent utilisés pour exprimer une émotion positive. Ils présentent également des différences : le rire contient des signes auditifs et visuels avec une durée relativement courte ; le sourire contient notamment des signes visuels avec une durée variable. Par contre, il y a des recherches qui se focalisent sur l'influence de déformation vocale du sourire pendant l'élocution.

Pour la détection dans un corpus réaliste, la différence entre le comportement spontané et délibéré du visage est étudiée par plusieurs de ces études [32], [142]. Dans [142], les auteurs affirment que les expressions spontanées sont différentes des expressions actées ou posés à la fois en apparence et en temps de réaction. L'étude [32] a montré que de nombreux types de sourires spontanés (par exemple, poli) ont une amplitude plus petite, une durée plus longue, et une évolution progressive (onset, offset) plus lente que les sourires posés. Cela signifie que les méthodes utilisées pour la reconnaissance de rire et sourire actée ou posée pourraient ne pas fonctionner correctement sur les expressions réalistes. Les méthodes utilisées dans la communauté pour la détection de rire et sourire sont présentées dans les sous-sections suivantes.

1.2.1 Détection de rire et sourire en vidéo

Au niveau de la détection de rire et de sourire en vidéo, comme pour les autres expressions faciales, les systèmes utilisant des modèles statistiques suivent l'algorithme présenté dans la Figure 2.

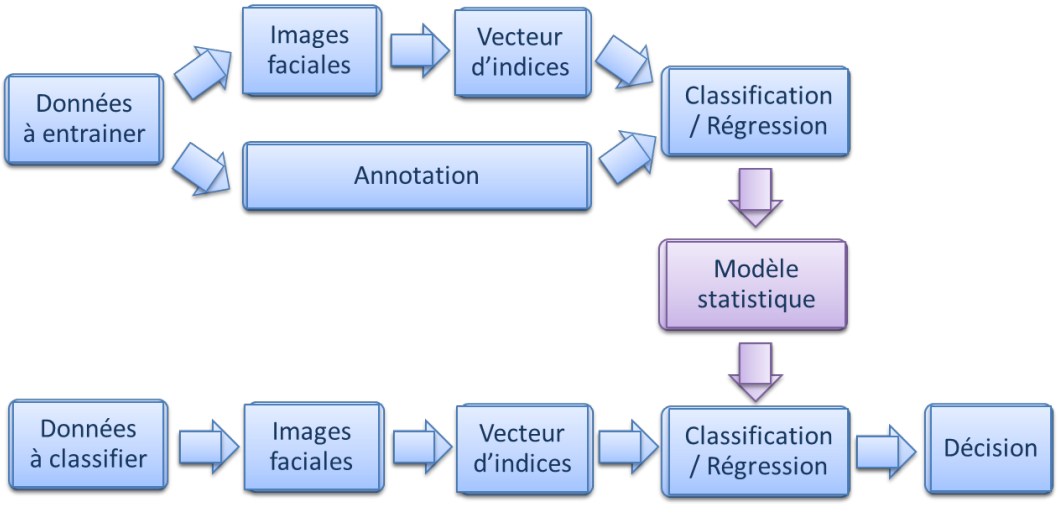

Figure 2 : Algorithme du système de la classification par modèle statistique

Il existe une variété de méthodes d'extraction d'indices dans la littérature telles que « Local Gabor Binary Patterns »[88], « Local Phase Quantization »[3], « Histogram

of oriented gradients » [41], « Haar filters » [76] et la détection des « Action Units » via le système du codage FACS [45]. Les méthodes peuvent être classées dans trois grandes catégories d'approches :

- Approches à base d'indices d'apparence (« appearance-based approach » en anglais) [10, 11, 5, 26, 55, 140], y compris principalement l'apparence locale et globale, qui étudient plutôt les textures locales ou globalement dans le visage causées par les expressions faciales.
- Approches à base d'indices géométriques (« geometric-feature-based approach » en anglaise) [25, 95, 96, 142], qui utilisent souvent des méthodes de suivi des points d'intérêt dans le visage [75, 131] (par exemple les points correspondant aux « Action Units » définis par Paul Ekman en 1978) pour étudier le contour du visage et de ses éléments afin d'adapter les modèles de l'expression tels que « Active Appearance Models » [44] et « Active Shape Models » [33].
- Mélange des 2 approches [78, 160]

De plus, beaucoup d'études se concentrent également sur des approches à base de modèles 3D [27, 29, 31, 115, 130, 152, 154, 155]. Avec la commercialisation de la caméra Kinect, la recherche de modèle 3D pour l'affect devient de plus en plus intéressante. Par contre, les approches à base de modèle 3D exigent souvent des données spécifiques comme l'image de profondeur ou une seconde caméra, etc. Ces approches ne peuvent pas être utilisées facilement avec la plupart des corpus existants.

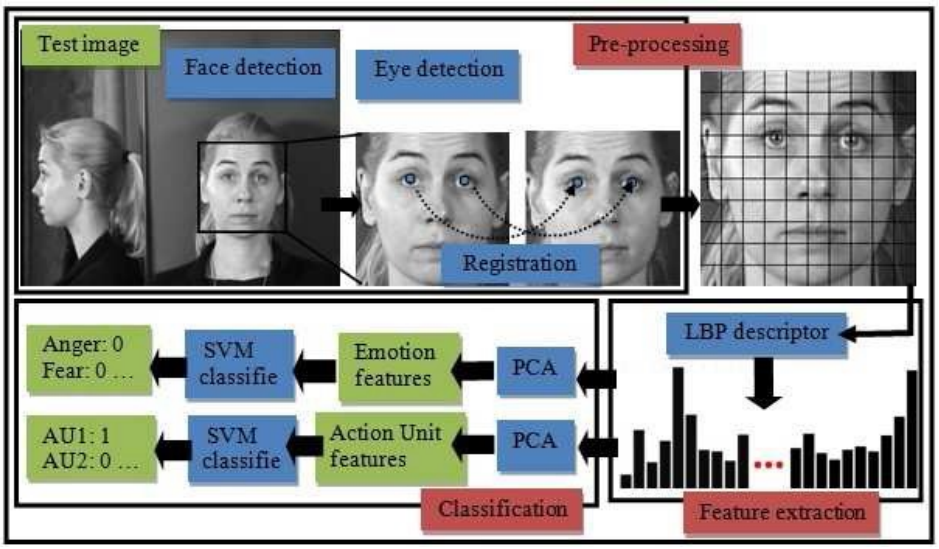

Figure 3 : Algorithme de la méthode LBP pour la classification statistique proposée par la conférence FERA11 [141] et AVEC 12 [112] comme la méthode de référence. (Source de l'image : [141])

En comparant les deux approches classiques, les approches à base d'indices géométriques semblent moins performantes selon les résultats des articles soumis à la

conférence FERA11[11], parce que les méthodes utilisant le suivi des points d'intérêt subissent plus d'erreurs de détection et sont moins robustes lors de l'initialisation. Ces problèmes seront fort probablement aggravés avec les données de personnes âgées. De plus, un avantage majeur de l'approche à base d'apparence est l'économie en coût de calcul.

Une méthode standard pour la détection des expressions faciales en utilisant l'approche à base d'indices globaux et locaux d'apparence est proposée par la conférence FERA11 [141] et AVEC 12 [112] comme méthode de référence. L'algorithme de la méthode est montré dans la Figure 3, elle suit les huit étapes ci-dessous :

1. Utiliser la détection de visage de Viola & Jones [145] pour récupérer l'image faciale
2. Utiliser la détection des yeux de Haar-cascade [98] pour l'échelle et la rotation planaire de la tête
3. Couper l'image faciale en 10x10 zones
4. Appliquer l'extracteur d'indice « uniform LBP » [93] avec un rayon de 1 pixel et 8 voisins (extension de la méthode « Local Binary Patterns » [92]) sur chaque zone afin d'obtenir l'histogramme LBP de chaque zone
5. Regrouper les histogrammes LBP des 10x10 zones dans un seul vecteur d'indices qui présente les informations de l'image faciale
6. Utiliser l'analyse en composantes principales [64] pour réduire la dimension des indices dans le vecteur entre les classes des émotions à détecter
7. Utiliser les indices dans le classifier SVM [35] pour entrainer le modèle et faire la classification
8. Détecter l'émotion de chaque trame dans une vidéo puis utiliser l'étiquette de l'émotion la plus présente pendant la durée du clip comme résultat de la détection.

A la suite du succès de la méthode LBP pour le traitement des images statiques, la méthode LGBP-TOP [4] (Local Gabor Binary Patterns from Three Orthogonal Planes) a été créée en combinant les méthodes utilisées par les gagnants de la conférence FERA11 dont la méthode de l'extraction d'indices LGBP [118, 119, 151] et la méthode d'extension de représentation des indices en volumes spatio-temporels TOP [62, 162]. Au lieu d'utiliser l'extracteur d'indices LBP sur l'image origine, la méthode LGBP applique la LBP sur l'image traitée par le filtre Gabor. Et pour la méthode d'extension TOP, au lieu d'utiliser les coordonnées horizontales et verticales pour représenter l'information statique d'un pixel dans une image fixe au format

[11] http://sspnet.eu/fera2011/

p(x,y), la méthode TOP représente également les informations dynamiques au niveau du temps (t), elle utilise trois plans pour la représentation des informations : plan x-y (information spatiale), x-t et y-t (informations dynamiques).

Par contre, les méthodes sont testées sur des corpus de sujets adultes avec des expressions actées ou posées. L'évaluation des méthodes sur un corpus réaliste d'interaction sociale entre les personnes âgées et un robot sera un défi. Pour la première expérience, un test sur une méthode standard, reconnue et conservative comme la méthode LBP qui est déjà testée sur de nombreux corpus sera plus conservative.

1.2.2 Détection de rire en audio

Au niveau de la détection de rire en audio, nous pouvons distinguer deux types de reconnaissance de rire: la reconnaissance avec une segmentation préalable ou la segmentation par la reconnaissance. Dans la première approche, des segments audio courts représentant des événements acoustiques (tout ce qui n'est pas un silence) sont manuellement ou automatiquement extraits d'un flux audio continu avant d'être classés. Quant à la seconde approche, la segmentation par la reconnaissance, il n'y a aucune connaissance préalable sur l'endroit où un événement acoustique commence et où il finit. Le flux audio entier est analysé pour distinguer les classes d'intérêt (par exemple la parole, des rires, silence, d'autres sons humains ou environnementaux, etc.).

Dans [69], les indices MFCC et les indices de « Modulation Spectrum » sont utilisés avec le classifieur SVM pour la détection de rire dans les salles de réunion. L'objectif principal des auteurs était la détection d'événements de rire, où plus d'une personne peuvent rire simultanément. Les indices spatiaux ont donc été calculés par la corrélation croisée des signaux audio acquis par deux microphones posés sur la table. Les objectifs de cette corrélation croisée sont de mieux distinguer le rire d'un participant et de multi-participants. Dans [136], de nombreux ensembles d'indices acoustiques (indice de codage de « Perceptual Linear Prediction » ou PLP, énergie, pitch et « Modulation Spectrum ») et des algorithmes de classification (« Gaussian Mixture Models » ou GMM, « Hidden Markov Models » ou HMM et « Multi Layer Perceptrons ») sont étudiés pour la classification entre le rire et la parole. La meilleure performance de référence est obtenue par les indices PLP avec le classifier GMM. Une amélioration a pu être observée en combinant les indices PLP avec le système GMM avec un système utilisant sur des indices de « pitch » avec le classifieur SVM. [110] envisage un problème de classification à 5 classes. Chaque segment audio est classé en quatre classes humaines (la respiration, le consentement, l'hésitation et le rire) et une classe fourre-tout utilisée pour modéliser le bruit de fond. La meilleure performance a été obtenue avec le classifieur HMM et les indices PLP.

Ces méthodes posent toutes le problème de la classification après la segmentation. Beaucoup d'autres travaux se concentrent sur la détection de rire en utilisant la segmentation par le schéma de classification. Dans [135], un flux est segmenté en intervalles parmi le rire, la voix et le silence à l'aide d'indices PLP et d'un classifieur GMM. Un décodeur Viterbi à trois états est d'abord utilisé pour trouver la séquence la plus probable des états à partir d'un flux. La séquence d'états est considérée comme une segmentation préliminaire. La probabilité de chaque segment à partir de chacun des modèles GMM est calculée pour prendre la décision de classification du segment. Dans [107], les indices MFCC et le classifieur HMM sont utilisés pour annoter un flux de données avec un ensemble de classes contenant des rires, de l'hésitation (« filler »), le silence et la parole. Un modèle de niveau supérieur, un modèle de langage bi-gramme, est utilisé pour modéliser explicitement l'ordre dans lequel les étiquettes apparaissent dans les données d'entrainement.

1.3 Détection de l'attention

En dehors des marqueurs affectifs, les marqueurs attentionnels sont également un indice performant pour mesurer la satisfaction et l'implication d'un sujet dans l'interaction humain-humain ou humain-robot.

La parole est un canal très important de l'interaction sociale et peut transporter une grande quantité d'informations sur l'intention et la motivation du locuteur. Une littérature abondante suggère également que le regard, l'orientation de la tête et l'orientation du corps jouent un rôle important lors de l'interaction sociale, et sont particulièrement utilisés et perçus comme un signal d'attention lors de l'interaction humaine [6, 129, 71, 106]. Vertegaal et al. [82] montre que le suivi de regard est un bon prédicteur de l'attention conversationnelle dans les conversations multi-locuteurs. En étudiant la relation entre le regard et la parole, ils montrent que les sujets regardent environ trois fois plus les personnes auxquelles ils parlent. Le travail de Maglio et al [144] montre que les gens ont tendance à regarder vers les sujets avec lesquels ils interagissent par la parole.

La détermination de l'orientation de la tête représente un domaine important de la recherche en interaction homme-machine (IHM). L'orientation de la tête de l'utilisateur présente une information riche dans une conversation homme-machine ou humain-humain (par exemple, pointer un objet, partager l'opinion par le hochement de tête, changer d'interlocuteur). Elle peut donc être la clé de nombreuses recherches dans le secteur IHM, tels que l'informatique affective, la segmentation des locuteurs dans le cas d'une conversation multi-locuteurs, le suivi de regard, etc. En considérant le problème de vue des personnes âgées et la difficulté de localisation des yeux dans la texture faciale des personnes âgées, la détection de la rotation de la tête nous semble une méthode performante pour notre expérimentation.

Des études physiologiques [74] ont démontré que la prédiction du regard de la personne provient d'une combinaison à la fois de l'orientation de la tête et de la direction du regard. En comparant le regard avec la détection de l'orientation de la tête, Johansson et al. [63] suggère que l'estimation du regard par l'orientation de la tête au lieu du suivi du regard a cependant l'avantage d'être plus robuste en ce qui concerne les mouvements de tête et le clignement de l'œil. Le travail précédent prouve également que l'orientation de la tête est un indicateur utile pour l'identification du destinataire en interaction multi-locuteurs [122], et qu'elle est suffisante pour déterminer la cible visuelle dans de nombreux cas, l'ajout du suivi du regard n'étant souvent pas nécessaires [68]. Il y a de nombreuses de recherches qui surveillent les sujets dans une « smart room » et utilisent l'orientation de leur tête pour mesurer leurs activités et leur focus visuel d'attention [12, 57, 85, 91, 99, 125, 133, 134, 137, 146, 147].

Dans [89], Erik Murphy-Chutorian résume les méthodes de détection automatique de la rotation de la tête apparues avant l'année 2009. Pour les recherches des années récentes, les recherches se sont concentrées sur le test des extracteurs d'indices et des différents modèles statistiques afin d'améliorer la performance de détection [60]. De plus, grâce à la commercialisation du Kinect de Microsoft, les approches utilisant la reconstruction du modèle 3D de la tête du sujet deviennent de plus en plus populaires [47, 72, 161].

Dans la communauté, certains des systèmes de détection de la rotation de la tête sont conçus pour les personnes âgées [124] mais peu d'entre eux sont effectivement évalués avec les personnes âgées [122]. En raison de la rareté de données de personnes âgées avec différentes orientations de la tête, une méthode non-statistique sera proposée en alternative dans mon expérimentation. En tout cas, le traitement d'image pour les personnes âgées subit la difficulté causée par les textures faciales des personnes âgées.

En dehors de la détection de rotation de la tête, une autre approche basée sur l'analyse du signal de parole peut être tentée pour distinguer si le sujet est en train de parler avec le robot ou à quelqu'un d'autre. Par exemple, en analysant un grand corpus de discours spontané, Campbell [21] a montré que la hauteur de la voix (« pitch » en anglais) F0 peut varier énormément quand la parole s'adresse aux enfants, aux membres de la famille ou à d'autres personnes auxquelles on s'adresse en général poliment.

1.4 Corpus existants

Dans ce sous-chapitre de ma thèse, des corpus reconnus dans la communauté de

l' « Affective computing » avec les études de la reconnaissance cognitive et affective correspondantes seront discutés. Les corpus avec une description brève sont montrés dans le Tableau 2.

Tableau 2 : Tableaux des corpus reconnus disponibles pour la communauté de l'étude de la reconnaissance affective et sociale.

Corpus	Tâche	Type	Taille
GENKI-4K	Rire et sourire	Spontané, images collectées sur internet	4000 images
Pointing04	Orientation de la tête	Acté, suivi visuel de marqueurs repérés sur les dimensions horizontales et verticales	15 sujets, 2790 images
CMU Multi-PIE	Orientation de la tête et émotion	Acté, capture d'image faciale sur 15 caméras avec différents angles horizontaux	337 sujet, >750000 image
JAFFE	Emotion	Acté, émotion basique	7 sujets, 213 images
DaFEx	Emotion	Acté, prononciation émotionnelle d'une phrase donnée et jeu de rôle	8 sujets, 1008 séquences vidéo
SEMAINE	Emotion et dimension abstraite	Spontané, conversation avec un avatar	150 sujets, 956 conversations
MMI	Emotion et AU	Acté, changement d'une émotion neutre à une émotion basique (Ekman) puis retour à l'émotion neutre	75 sujets, 2900 séquences vidéo
Cohn-Kanade	Emotion et AU	Acté, changement de l'émotion neutre à une émotion basique (Ekman)	97 sujets, 486 séquences vidéo
GEMEP	Emotion et AU	Acté, par des professionnels avec mise en contexte	10 sujet, >7000 séquences vidéo

La détection de l'orientation de la tête avec la précision au niveau de l'angle exige des images annotées avec un label de l'angle précis de l'orientation en horizontale et parfois également en verticale. Le corpus Pointing04 [53] et le corpus CMU Multi-PIE [121] sont souvent utilisés dans la communauté. L'inconvénient du corpus CMU Multi-PIE est qu'il ne capture pas la rotation de la tête dans la condition nature, mais elle déplace les caméras sur le profil du sujet pour simuler une rotation de la tête. De plus, le corpus CMU Multi-PIE contient notamment les images d'orientation horizontale.

Pour la reconnaissance affective, il existe de nombreuses bases de données d'expressions du visage actées ou posées en vidéo telles que le corpus Cohn-Kanade [65], le corpus GEMEP [9], le corpus d'expression facial MMI [97, 139], le corpus DaFEx [14] et le corpus JAFFE [79]. Mais peu de corpus réalistes des expressions spontanées comme le corpus GENKI-4K [138] et le corpus SEMAINE [86] existent. Ils sont encore moins nombreux en ce qui concerne des données réalistes avec des

personnes âgées. La plupart des chercheurs ont testé et validé leurs méthodes de détection sur les corpus actés ou posés [38, 59, 73, 120].

De plus en plus de recherches se réorientent sur l'analyse automatique des données d'expressions spontanées du visage [10, 11, 30, 31, 32, 58, 77, 78, 115, 142, 155]. Certains d'entre eux étudient la reconnaissance automatique des AU plutôt que les textures des émotions faciales spontanées [10, 11, 30, 31, 142]. La différence entre le comportement spontané et délibéré du visage est étudiée par plusieurs de ces études [32], [142]. L'étude [32] a montré que de nombreux types de sourires spontanés (par exemple, poli) ont une amplitude plus petite, une durée plus longue, et une évolution progressive (onset, offset) plus lente que les sourires posés. En outre, l'étude [142] a montré que les actions spontanées de sourcil (AU1, AU2 et AU4 dans le système FACS) ont des caractéristiques morphologiques et temporelles différentes (intensité, la durée et l'ordre d'apparition) des actions de sourcils posées.

Pour le workshop « Emotion Recognition In The Wild Challenge 2014» (EmotiW 2014)[43], « in the wild » signifie prenant en compte les variabilités de l'environnement, de la condition de la luminosité, de la pose de la tête, etc. De plus, il utilise un corpus de données actées et segmentées de l'AFEW [42] comme corpus de référence pour évaluer les systèmes de reconnaissance automatique. Dans notre expérimentation, nous considérons une évaluation « in the wild » comme l'évaluation en flux continu dans un corpus réaliste non segmenté.

Le corpus des personnes âgées, collecté dans le cadre du projet ROMEO2 permet de mener de nombreuses études sur les dimensions affectives et sociales en interaction avec un robot. Il est décrit dans la partie II « Collecte, annotation et analyse du corpus ROMEO2 » de la thèse. Il m'a permis d'étudier la détection des marqueurs affectifs et attentionnels dans un corpus « real-life » avec une approche par frame, segment et « in the wild ».

2 Techniques d'analyse de l'image

Etant dans un groupe expert sur la détection des émotions dans l'audio, mon sujet a été complémentaire sur la partie visuelle afin de fusionner les canaux audio et visuel. Dans ce deuxième chapitre de ma thèse, un ensemble de techniques d'analyse d'image qui couvrent la détection de visage, l'extraction d'indices faciaux, la classification, la fusion multimodale et l'évaluation de performance du système que j'ai utilisé dans mon expérimentation sera présenté.

2.1 Détecteur de visage utilisant Haar Cascades

En tant que première étape du traitement d'image faciale, la méthode de détection de visage de Haar-Cascade permet de trouver et localiser des visages très efficacement. Cette méthode de détection d'objets à l'aide de plusieurs classificateurs Haar en cascade est proposée par Paul Viola et Michael Jones dans leur article [145] en 2001. C'est une approche à base d'apprentissage automatique où une fonction cascade est formée à partir d'un grand nombre d'images positives et négatives. Le classifieur est ensuite utilisé pour détecter des objets dans d'autres images.

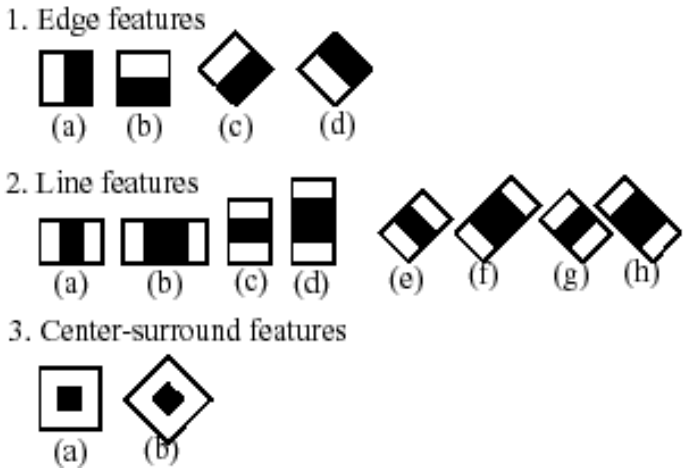

Figure 4 : Exemple des indices Haar. (Source de l'image : http://docs.opencv.org)

Les indices utilisés pour la classification sont ceux de Haar montrés dans Figure 4, ils sont similaires à des noyaux de convolution. Chaque indice est une valeur unique obtenue en soustrayant la somme de pixels dans la zone blanche de la somme des pixels dans la zone noire. Pour simplifier le calcul de la somme de pixels dans un rectangle, les auteurs introduisent une méthode qui s'appelle « integral images » qui pré-calcule le somme des pixels dans le rectangle $S_{\{0,0,x,y\}}$ pour chaque coordonnées {x,y}, le somme des pixels de n'importe quel rectangle $S_{\{a,b,c,d\}}$ peut donc être calculé simplement par $S_{\{a,b,c,d\}}=S_{\{0,0,c,d\}}+S_{\{0,0,a,b\}}-S_{\{0,0,c,b\}}-S_{\{0,0,a,d\}}$.

Même si le coût de calcul de chaque indice est ainsi réduit de manière importante, le nombre d'indice est trop élevé si on prend tous les indices de Haar dans une image, par exemple 160.000 indices pour une image de 24x24 pixels. Les auteurs utilisent la méthode « Adaboost » pour ajuster le poids de vote de chaque indice et réduire le nombre des indices à un peu plus de 6.000. Selon l'article, le ratio de bonne reconnaissance atteint 95% même avec 200 indices.

Le calcul de tous les 6.000 indices pour chaque fenêtre dans une image est toujours coûteux. Les auteurs utilisent une conception en cascade des classifieurs. Le mot "cascade" associé au nom de classificateur signifie que le classificateur résultant se compose de plusieurs classificateurs simples (étapes) qui sont appliqués par la suite à une région d'intérêt jusqu'à ce que, à un moment soit le candidat est rejeté, soit toutes les étapes sont passés. La méthode utilisée regroupe plus de 6.000 indices en 38 étapes avec 1, 10, 25, 25 et 50 indices au cours des cinq premières étapes. A chaque étape, les indices sont appliqués un par un. Si une fenêtre échoue la première étape, le système ne considère pas les indices restant sur celle-ci. Par contre si elle passe, la deuxième étape des indices sera appliquée et le processus se poursuivra. La fenêtre qui passe toutes les étapes est désormais considérée comme une région de visage.

Grâce à son efficacité et sa performance, l'ensemble de cette méthode de détection d'objet est largement utilisée dans le domaine de traitement d'image et particulièrement dans la détection et le suivi de visage, la détection et le suivi des yeux, etc. La méthode est disponible et embarquée dans la librairie OpenCV.

2.2 « Local Binary Patterns »

L'opérateur « Local Binary Pattern » (LBP) est un descripteur simple mais puissant pour l'analyse de texture [92]. Il nécessite une puissance de calcul limitée, donc est idéal pour les applications en temps réel. Sa robustesse au changement monotone d'échelle de gris convient à des applications telles que l'analyse de l'image du visage où les variations d'éclairage peuvent avoir des effets majeurs sur l'apparence.

Le processus du calcul des LBP suit les étapes suivantes :

1. Diviser l'image en blocs de pixels (souvent 10x10).
2. Pour chaque pixel dans un bloc, comparer la valeur de pixel avec celle de ses voisins. La taille du voisinage est contrôlable, en général, on utilise les pixels du voisin avec une distance d'un pixel, donc le nombre de voisins est de 8. La comparaison est effectuée soit dans le sens horaire soit antihoraire.
3. Si la valeur du pixel du voisin est supérieure à la valeur du pixel au centre, un code binaire "1" est généré sinon un "0". Pour un voisinage de 3x3, un indice de huit bits sera généré et peut être converti en valeur décimale. L'ensemble du processus de l'étape 2 et 3 est illustré dans la Figure 5.

4. Un histogramme de la répartition du niveau de gris des pixels est calculé pour le bloc.
5. Les histogrammes pour tous les blocs de l'image sont enchaînés pour obtenir un seul vecteur, donc le descripteur final de l'image.

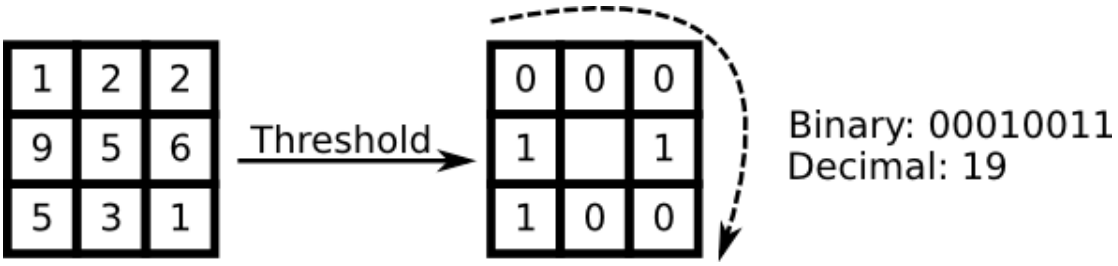

Figure 5 : Calcul de la valeur LBP pour un pixel (le pixel au centre) avec les 8 pixels voisins autour. (Source de l'image : http://docs.opencv.org)

Beaucoup de variations sur l'opérateur LBP origine ont été proposées. Une extension permet à l'opérateur LBP d'utiliser différentes tailles de voisinage en pixels [93]. Une autre modification introduite dans [93] est le LBP uniforme (« Uniform-LBP » en anglais) qui peut être utilisé pour réduire la longueur du vecteur des indices à environ 22% de l'origine en regroupant les indices qui ont moins de 1 changement binaire entre le code 0 et 1 dans un seul indice.

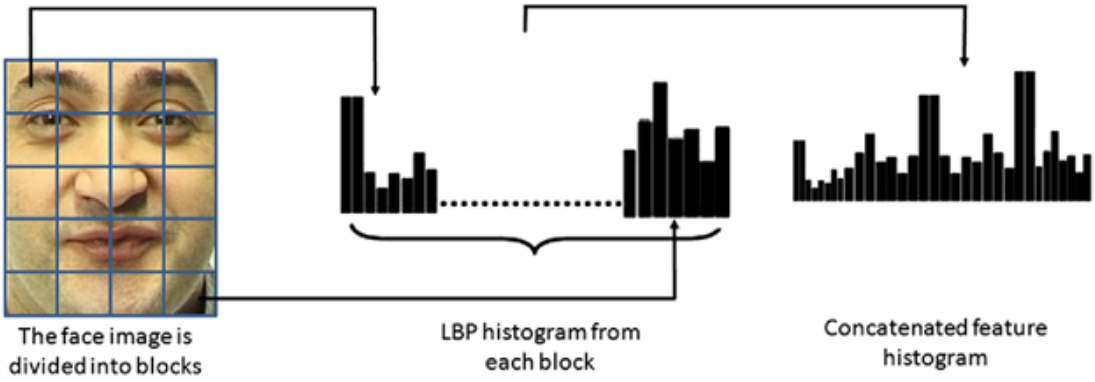

Figure 6 : Extraction de vecteur de l'histogramme de LBP pour une image. (Source de l'image : http://what-when-how.com/face-recognition/)

2.3 « Support Vector Machines »

Le classifieur de machines à vecteurs de support (« Support Vector Machines » en anglais) est très largement utilisé dans le domaine informatique. À l'origine, le SVM est une technique pour la construction d'un classificateur binaire (2 classes) optimal. La classification multi-classes par les SVM est possible en utilisant plusieurs SVM à 2 classes. Plus tard, la technique a été étendue pour résoudre des problèmes de régression et de regroupement.

L'idée de base de l'algorithme de SVM est d'utiliser la distance scalaire entre les vecteurs pour chercher un hyperplan optimal qui permet de séparer les points appartenant à différentes classes. Comme montré dans la Figure 7, l'hyperplan optimal est celui qui a la distance maximale entre les points de données des deux classes, c'est-à-dire l'hyperplan avec la plus grande marge entre les deux espaces. Afin de simplifier la séparation des espaces de très grande dimension, les points sont projetés dans un nouvel espace en utilisant une fonction à noyau. De ce fait, les SVM sont particulièrement adaptés à des problèmes de classification dans des espaces complexes.

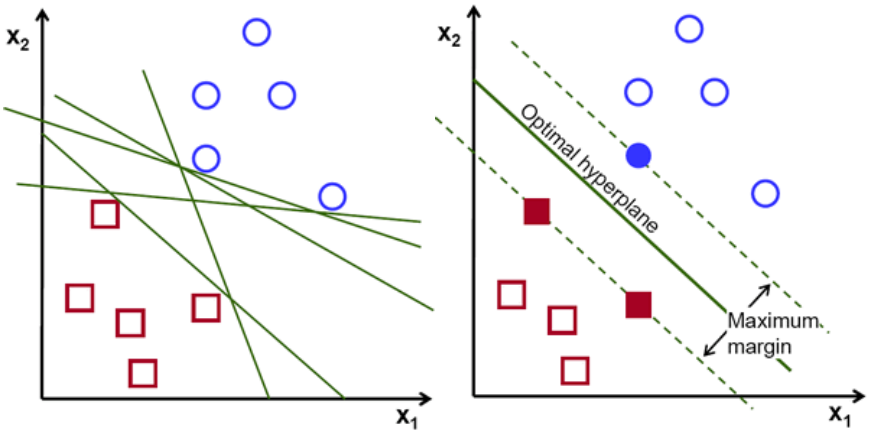

Figure 7 : Exemples des hyperplans possibles (gauche) et de l'hyperplan optimal (droit) pour la classification des deux classes dans une espace 2D. (Source de l'image : http://docs.opencv.org)

Le noyau le plus populaire utilisé avec SVM est le noyau de base radiale, il s'agit d'un bon choix de noyau dans la plupart des cas et il est représenté par l'équation suivante:

$$K(x_i, x_j) = e^{-\gamma||x_i - x_j||^2}, \ \gamma > 0$$

L'implémentation du classifieur dans OpenCV utilise la librairie LibSVM [24]. En dehors du classifieur SVM, d'autres classifieurs non-linéaires étaient également envisageables comme l'apprentissage avec multi-noyau [7], les « Relevance Vector Machines » [132], le « deep-learning neural network » [15].

2.4 Fusion multimodale et cascade

La plupart des études de reconnaissances affectives audiovisuelles existantes se focalisent sur les émotions basiques d'Ekman, mais il existe également quelques études pour détecter les émotions sociales, comme les travaux de Zeng et al. [157, 158, 159], qui a ajouté quatre états cognitifs (l'intérêt, la perplexité, la frustration et l'ennui) considérant l'importance de ces états cognitifs dans l'interaction humain-machine. De plus, les états « intérêt » et « ennui » sont fortement liés à l'attention du locuteur. Il

existe également quelques d'études sur la reconnaissance audiovisuelle des émotions spontanées [22, 36, 156], parmi lesquelles Fragopanagos et Taylor [36] et Caridakis et al. [22] utilisent les données recueillies dans les scénarios de Magicien d'Oz (« Wizard of Oz » en anglais). De plus, d'autres auteurs ont étudié la fusion des informations à partir de l'expression faciale et du mouvement de la tête [31, 61, 160] ou la fusion de l'expression du visage et du mouvement du corps [8, 54, 66] dans le but d'améliorer la performance de reconnaissance affective.

Au niveau de la fusion multimodale pour la reconnaissance affective, il y a principalement deux types de fusions : la fusion précoce (« Early fusion » en anglais) et la fusion tardive (« Late fusion » en anglais). La fusion précoce se focalise sur la fusion au niveau des indices multimodaux ; différents classifieurs sont utilisés dans les études publiées dans la communauté, par exemple, les HMMs [123, 157, 159], les réseaux de neurones artificiels [36, 67], les réseaux bayésiens [114]. La difficulté de la fusion précoce concerne la synchronisation entre les indices de différentes modalités au niveau de l'échelle temporelle et unitaire ainsi que l'augmentation du nombre dimension des indices à calculer. La fusion tardive se focalise sur la fusion au niveau des résultats des modalités. La plupart des études en reconnaissance affective audio-visuelle utilisent la fusion tardive [20, 51, 56, 148, 158], qui analyse indépendamment les expressions sur le canal auditif et visuel puis combinent les résultats de la reconnaissance uni-modale à la fin par des stratégies différentes, par exemple par le poids de vote [56] ou par règles [100].

En conclusion, il y a plusieurs choses à considérer pour le choix de la fusion dans la détection audio-visuelle:

- Synchronisation entre les modalités : le temps nécessaire de reconnaissance pour chaque modalité n'est pas identique
- Dépendance entre les modalités : par exemple la reconnaissance vocale liée au mouvement de bouche [18]
- Facilité d'implémentation : compromis entre le temps de traitement et la performance
- Performance de chaque modalité : la robustesse de chaque émotion détectée n'est pas au même niveau
- Contrainte de l'environnement : bruit, distance, luminosité, aberration chromatique, etc.

Pour la première expérience, la fusion tardive nous semble un bon choix pour faciliter la tâche et avoir une première idée de performance de la fusion. Pour la fusion tardive audio-visuelle, une fusion en cascade peut être très utile dans le cas où la classification d'une modalité est plus « experte » que d'autres pour certains problèmes, par exemple, une détection de la rotation de la tête vers une autre direction peut signifier la perte de l'attention temporelle sur l'interaction, et la détection de la voix

en audio peut efficacement permettre de distinguer une bouche parlante d'une bouche souriante.

2.5 Evaluation de performance

Les méthodes de reconnaissance automatique exigent des normes standards pour évaluer leur efficacité. Les résultats pour un classifieur binaire (deux classes ou une classe contre toutes les autres classes) peuvent être représentés dans une matrice de confusion dans la forme du Tableau 3, où les quatre mesures doivent être calculés : le nombre d'entités correctement reconnues comme appartenant à une classe (Vrais Positifs VP), le nombre d'entités correctement reconnues comme n'appartenant pas à une classe (Vrais Négatifs VN), le nombre d'entités incorrectement reconnues comme appartenant à une classe (Faux Positifs FP), et le nombre d'entités incorrectement reconnues comme n'appartenant pas à une classe (Faux Négatifs FN).

Tableau 3 : Matrice de Confusion pour l'évaluation des modèles de reconnaissance automatique. VP signifie les vrais positifs, VN signifie les vrais négatifs, FP signifie les faux positifs, FN signifie les faux négatifs.

		Prédiction	
		Positive	Négative
Entrée	Positive	VP	FN
	Négative	FP	VN

Afin de comparer la performance des systèmes de classification des différentes situations, plusieurs types de scores peuvent être calculés à partir de la matrice de confusion :

Taux de bonne reconnaissance et taux d'erreur

Le taux de bonne reconnaissance (« accuracy » en anglais) correspond au rapport des entités correctement classées, il peut être calculé par :

$$R_A = \frac{VP + VN}{VP + VN + FP + FN}$$

Il existe également le taux d'erreur qui correspond au rapport des entités mal classées :

$$R_E = \frac{FP + FN}{VP + VN + FP + FN}$$

Ces deux mesures nous permettent d'avoir une idée globale de la performance du système, par contre, elles ne présentent pas d'informations avancées sur les erreurs commises entre les faux positifs et négatifs.

Précision, rappel et F-mesure

La précision correspond au rapport des entités d'entrée positive parmi les entités classés en positif par le système, elle peut être calculée par :

$$P = \frac{VP}{VP + FP}$$

Le rappel correspond au rapport des entités classé en positive parmi les entités d'entrée positive, elle peut être calculée par :

$$R = \frac{VP}{VP + FN}$$

Le F-mesure combine les mesures de précision et de rappel pour présenter la performance du système en fonction de :

$$F = \frac{2 * F * R}{F + R}$$

A noter que, dans les mesures de précision, de rappel et de F-mesure, seules les performances du système sur la classe positive sont évaluées. Dans le cas où les nombres d'entités positives et négatives ne sont pas balancés, ces mesures ne peuvent pas présenter correctement la performance complète du système. Par exemple, dans le cas où il y a peu d'entrés positives par rapport à négatives, la précision du système sera sous-estimée.

Taux d'erreur balancé

Dans le cas où les nombres d'entrées positives et négatives ne sont pas balancés, le taux d'erreur balancé (« Balanced Error Rate » en anglais) est souvent utilisé pour présenter la performance du système en calculant la moyenne des erreurs de chaque classe :

$$BER = 0{,}5 * (\frac{FN}{VP + FN} + \frac{FP}{FP + VN})$$

Coefficient de corrélation linéaire de Bravais-Pearson et valeur p

Pour mesurer la relation entre deux variables continues de même longueur x et y, la corrélation linéaire de Bravais-Pearson (« Pearson product-moment correlation » en anglais) est souvent utilisée pour évaluer la force de la relation linéaire entre x et y. La fonction est écrite par :

$$Cor(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Où $Cov(x,y)$ désigne la covariance des variables x et y; σ_x et σ_y pour leurs écarts types.

Le coefficient de corrélation est compris entre -1 et 1, plus le coefficient est proche des valeurs extrêmes -1 et 1, plus la corrélation entre les variables est forte.

Dans un test statistique, la valeur p (« p-value » en anglais) est la probabilité d'obtenir la même valeur (ou une valeur encore plus extrême) du test si l'hypothèse nulle était vraie. La valeur p est souvent utilisée pour présenter le niveau significatif (significativité) statistique d'une hypothèse. Les seuils suivants sont utilisés en général :

- $p < 0,01$: très forte présomption contre l'hypothèse nulle
- $0,01 < p < 0,05$: forte présomption contre l'hypothèse nulle
- $0,05 < p < 0,1$: faible présomption contre l'hypothèse nulle
- $p > 0,1$: pas de présomption contre l'hypothèse nulle

Une valeur p supérieure à 0,05 ne permet pas de conclure de manière décisive.

Evaluation dépendante de la tâche

A noter que, l'évaluation de la performance doit considérer également la nature de corpus de test et le type d'annotations appliquées. Un corpus réaliste pose souvent plus de difficulté, donc un score relativement faible. Et l'évaluation utilisant une annotation en temps continu entraine souvent une chute de performance par rapport à une évaluation du système sur les segments entiers. De plus, certaines tâches spécifiques ajoutent toujours des défis supplémentaires pour un système, par exemple, la reconnaissance de visage pour les personnes âgées entraîne plus d'erreur que pour la population plus jeune.

Synthèse de l'État de l'Art

Dans la première partie de ma thèse, nous avons présenté les notions du domaine de l'« Affective computing » ainsi que les travaux existants reconnus dans la communauté sur la détection visuelle et audio-visuelle des émotions et des marqueurs affectifs et d'attention. Nous nous sommes focalisés sur la détection des marqueurs affectifs « Affect bursts » particulièrement le rire et le sourire comme une méthode complémentaire. Les méthodes de détection de rire et sourire en vidéo ont été présentées plus en détail dans le corps de la partie. Selon les différences entre les expressions spontanées et les expressions actées ou posées au niveau de l'apparence et du temps de réaction montrées dans [142], les méthodes validées par des données actées ou posées pourraient ne pas fonctionner correctement sur les expressions réalistes, particulièrement pour les personnes âgées. Donc pour notre expérience, un test sur une méthode standard, reconnue et conservative comme la méthode LBP pour la détection visuelle qui est déjà testée sur de nombreux corpus sera plus conservative.

En dehors des marqueurs affectifs, les marqueurs attentionnels sont également un indice performant pour mesurer la satisfaction et l'implication d'un sujet dans l'interaction humain-humain ou humain-robot. Plusieurs méthodes de détection visuelle de l'attention sont présentées dans cette partie, en considérant le problème de vue des personnes âgées et la difficulté de localisation des yeux dans la texture faciale des personnes âgées, la détection de la rotation de la tête nous semble une méthode performante pour notre expérimentation. En dehors de la détection de rotation de la tête, une autre approche basée sur l'analyse du signal de parole peut être tentée pour distinguer si le sujet est en train de parler avec le robot ou à quelqu'un d'autre.

Je me suis concentré sur le traitement de l'image pour la reconnaissance automatique dans notre expérimentation. Un ensemble de techniques d'analyse d'image qui couvre la détection de visage, l'extraction d'indices faciaux, la classification, la fusion multimodale et l'évaluation de performance du système est également présenté dans le deuxième chapitre de cette partie de ma thèse. Au niveau de la fusion multimodale, les avantages et les inconvénients entre les deux grandes catégories de fusion (la fusion précoce et tardive) ainsi que les travaux correspondants sont listés pour comparaison. De plus, un ensemble de considération pour le choix du type de fusion pour notre expérimentation est également présenté. Pour la première expérience, la fusion tardive nous semble un bon choix pour faciliter la tâche et avoir une première idée de la performance de la fusion. Pour la fusion tardive audio-visuelle, une fusion en cascade peut être très utile dans le cas où la classification d'une modalité est plus « experte » que d'autres pour certains problèmes, par exemple, une détection de la rotation de la tête vers une autre direction peut signaler la perte de l'attention temporelle sur l'interaction, et la détection de voix en audio peut efficacement distinguer une bouche parlante d'une bouche souriante.

De plus, un résumé des corpus existants dans le domaine « Affective Computing » est également fourni. Bien que la plupart des corpus dans la communauté soient posés ou actés, de plus en plus de recherche se réorientent vers l'analyse automatique des données d'expressions spontanées du visage. Le manque de données ciblées sur les personnes très âgées exige une collection de données bien organisée et bien annotée pour satisfaire le besoin de recherche en « affective computing ».

Partie II: Collecte, annotation et analyse du corpus ROMEO2

Pour comprendre efficacement et modéliser le motif du comportement des personnes très âgées en présence d'un robot, des données pertinentes sont nécessaires. Ce type de données n'est cependant pas facile à obtenir et à partager en raison de nombreux facteurs. Les corpus de données enregistrées avec les jeunes pourraient être plus faciles à créer, mais ils ne répondent pas aux exigences des études sur lesquelles je me suis concentré dans ma thèse.

Cette partie de ma thèse présente le corpus réaliste ROMEO2 portant sur l'interaction sociale entre des personnes âgées et le robot humanoïde Nao collecté dans deux maisons de retraite. La collecte de données est une partie du projet français ROMEO2[12] (2013-17) qui suit le projet ROMEO [19, 37]. Le but du projet est de développer un robot humanoïde qui peut agir comme un assistant d'accompagnement pour les personnes souffrant de perte d'autonomie. Dans cette perspective, le robot est en mesure d'aider une personne dans ses tâches quotidiennes quand elle est seule. Le but de cette étude est de concevoir un système interactif affectif entraîné avec des marqueurs interactionnels, émotionnels et de la personnalité. Mon travail a porté sur la détection du rire et sourire et des marques d'attention dans ce projet.

Depuis le succès d'Eliza [149], la plupart des robots de conversation ont suivi les mêmes principes pour améliorer leur score au test de Turing. Le test de Turing est utilisé comme un critère de l'intelligence d'un programme d'ordinateur pour évaluer la capacité du programme à remplacer un agent humain dans une conversation en temps réel avec un humain. Il induit les utilisateurs qu'ils ne puissent pas se rendre compte qu'ils ont effectivement parlé à une machine. L'idée de base est la reconnaissance de mots clés ou de phrases dans l'entrée du sujet humain et l'utilisation efficace de ceux-ci (mots ou phrases) dans les réponses préalablement préparées ou prédéfinies afin de continuer la conversation d'une manière qui semble avoir sens pour l'humain. Par exemple, lorsqu'une entrée contient les mots «mère» ou «fils», la réponse du programme est généralement «Dites-moi plus sur votre famille » [149].

Notre dialogue humain-robot a été conçu dans le même esprit qu'Eliza. Les principaux défis sont de donner à la conversation un assez bon niveau de sens et de faire focaliser le sujet âgé dans la conversation autant que possible. Cependant, contrairement à Eliza et aux robots de conversation en général, comme le robot est dans la même

[12] http://projetromeo.com

pièce que la personne et donc visible pour elle, nous nous sommes concentrés sur le fait que le robot devrait être considéré comme un appareil intelligent, pas un humain.

Les objectifs initiaux sont dont

- Obtenir un premier retour des personnes âgées
- Valider et améliorer les scénarios envisagés
- Obtenir un corpus d'interaction entre les personnes âgées pour les recherches correspondantes
    - Reconnaissance des émotions et de rires à partir de l'audio [128]
    - Détection multimodale de rire et de sourire [153]
    - Détection de comportement social – attention [117]

Au niveau de ma contribution, j'ai participé à la deuxième session de la collection du corpus, conçu le schéma d'annotation audio et vidéo avec les collègues, annoté une partie du corpus, relu et corrigé les annotations des autres personnes, complété les résultats d'analyse des corrélations inter-questionnaires, et réalisé les analyses sur les corrélations inter-annotations et questionnaire-annotation. La collecte du corpus est un travail important qui a donc impliqué plusieurs personnes de l'équipe sous la coordination de ma directrice de thèse Laurence DEVILLERS et j'y ai naturellement participé.

Dans cette partie de ma thèse, le protocole de collecte de données et les scénarios d'interaction conçus ainsi que le corpus recueilli (27 sujets, âge moyen: 85) seront présentés dans le chapitre 3. Nous décrirons ensuite le schéma d'annotation et la quantité d'évènements annotés dans le chapitre 3.9. Nous discuterons des résultats obtenus à partir de l'analyse des annotations et de deux questionnaires (un questionnaire de satisfaction, et le Big-five qui est un questionnaire sur la personnalité) dans le chapitre 4.

3 Collecte

3.1 Données ciblées

Pour mener efficacement une étude sur l'interaction sociale des personnes âgées avec un robot, il est nécessaire de collecter des données en situation. En fait, ce type de données est plutôt rare et ne peut être ni collecté dans un laboratoire ni dans les émissions de télévision ou les conversations téléphoniques [23]. En outre, en raison du problème éthique et linguistique, ce type de contenu ne peut pas être facilement partagé avec d'autres chercheurs. Le corpus pertinent doit, de notre perspective, présenter des sujets âgés en conversation spontanée avec le robot.

Pour répondre à ces exigences, notre stratégie était de chercher la population ciblée dans une maison de retraite de personnes âgées, de concevoir plusieurs scénarios de conversation intéressants de vie de tous les jours qui devaient encourager les gens à coopérer avec le robot, et d'utiliser un schéma de Magicien d'Oz (« Wizard of Oz » en anglais) pour contrôler le robot afin de lui permettre d'adapter précisément et rapidement son comportement à la plupart des situations. Les maisons de retraite sont deux EHPADs français (hébergement public de personne-âgée non-autonome) à Montpellier. La collecte du corpus a été faite en collaboration entre notre laboratoire LIMSI et l'association Approche [116], également partenaire du projet ROMEO2 Nous nous sommes également concentrés sur le fait que le robot doit être considéré comme une machine intelligente, non pas un être humain. Ainsi, de nombreuses phrases sont utilisées par le robot Nao pour souligner cela, comme « Je viens de sortir de ma boîte », « Je viens de quitter l'usine », « J'ai beaucoup d'amis robots » ou « Je dois recharger ma batterie ».

3.2 Scénarios d'interaction

La conversation est subdivisée en plusieurs scénarios indépendants qui doivent être exécutées dans un ordre spécifique. La Figure 8 représente un exemple de l'interaction sociale entre une personne âgée et le robot Nao.

Les scénarios d'interactions sociales étaient:

- Introduction
- Stimulation de mémoire de court-terme: repas, médicament
- Interaction sociale: appeler un parent
- Stimulation cognitive: jeu de reconnaissance de chansons

Figure 8 : Exemple d'une interaction sociale entre une personne âgée et le robot Nao

Dans le premier scénario (Introduction), le robot Nao se présente et annonce ses capacités (à savoir qu'il peut parler, chanter et se déplacer) pour susciter la curiosité de la personne et lui donner envie de parler. Il pose ensuite au sujet des questions personnelles qui incluent son nom, son âge, sa durée résidante dans cet établissement, la composition de sa famille et ainsi de suite. Dans le second scénario, Nao essaie de faire parler le sujet plus fréquemment dans une conversation sociale telle que la météo du jour et leurs jeux favoris. Dans ce scénario, NAO stimule le mémoire à court-terme des sujets en leur demandant le menu de leur repas précédent et s'ils avaient des médicaments à prendre. Dans le troisième scénario, Nao essaie de parler de la famille et des enfants pour encourager la personne à appeler un parent. Dans le dernier scénario, l'objectif principal est de stimuler cognitivement la personne. Nao tente d'identifier l'intérêt du sujet parmi les films, la cuisine et le programme TV. Il joue alors une trentaine de secondes de vieilles chansons françaises et demande au sujet s'il a reconnu le titre de la chanson ou le nom de l'interprète.

3.3 Magicien d'Oz

Par l'utilisation de la méthode Magicien d'Oz, Nao est contrôlé à distance par un expérimentateur qui observe la conversion avec le sujet et réagit en conséquence. Le contenu de chaque scénario est défini et Nao (lequel est contrôlé par l'expérimentateur) obéit à un schéma d'arbre de conversation pour exécuter la prochaine action (lire un texte, lancer une chanson ou faire un geste). En outre, grâce à la facilité d'utilisation de la fonction de synthèse vocale à partir de texte de Nao, l'expérimentateur peut dynamiquement taper et envoyer un texte court comme le nom de la personne. Lors de situations difficiles lorsque la personne insiste ou ne suit pas le schéma de l'arbre de la conversation, l'assistant peut utiliser les 46 phrases génériques telles que « il est vrai», « Oui », etc. Le nombre moyen de phrases par session était de 82 phrases. Le nombre de phrases différentes du WOZ est de 265 (y

compris les phrases génériques) avec beaucoup de phrases empathiques telles que « J'aime bien ton prénom ».

L'objectif principal de la méthode du Magicien d'Oz (« Wizard of Oz » en anglais) est de profiter des capacités de communication de NAO et de construire une interaction sociale entre le robot et les personnes âgées. Par conséquent, l'outil que nous avons utilisé consiste en un logiciel avec une interface graphique et est globalement conçu pour envoyer les commandes à NAO comme les déclarations de texte, les gestes et les sons de jeu (par exemple les vieilles chansons). Dans le but d'optimiser la spontanéité et la rapidité de réaction de NAO, presque tous énoncés de parole sont codés à l'avance. En outre, l'expérimentateur peut mettre à jour dynamiquement quelques champs de texte (par exemple le nom de la personne) ou ajouter un texte libre afin de maintenir la conversation en correspondance avec le thème actuel du sujet quand il ne suit pas le scénario. Pour rendre l'utilisation de texte libre aussi limitée que possible, de nombreuses expressions génériques (par exemple, «Oui», «Non», «Je vois », « Vous m'entendez », « Je suis désolé », etc.) sont disponibles pour l'expérimentateur. Chaque scène dans un scénario est construite par un schéma d'arbre de dialogue. Selon l'action et la réaction du sujet, l'expérimentateur doit choisir le nœud suivant du dialogue à utiliser.

3.4 Equipement

Pour cette collecte de données, nous nous sommes focalisés sur les deux modalités: audio et vidéo. Un fichier journal est également disponible pour chaque conversation. Il contient tous les informations temporelles liées aux actions de NAO et peut être utilisé pour reconstruire le schéma de la conversation. En outre, il peut être également utilisé pour extraire des informations utiles telles des déclarations répétées et le temps de réaction du part du sujet, etc.

Figure 9 : Exemple des points de vue des caméras pendant la collecte du corpus ROMEO2. (Image gauche : prise par la caméra frontale, image droite : prise par la caméra de profil)

En dehors de la caméra vidéo de Nao et des 4 microphones embarqués, nous avons également utilisé une caméra du type webcam HD pour capturer l'expression du visage (un écran blanc a été mis en place derrière la personne), une caméra HD pour

enregistrer l'ensemble de l'interaction en vision de profil et une micro lavallière pour obtenir un enregistrement audio isolé de haute qualité. Un exemple des points de vue des caméras est montré dans la Figure 9.

3.5 Participant

Le nombre de sujets est de 27 personnes (4 hommes et 23 femmes) enregistrées lors de deux séances (14 sujets en novembre 2013 et 13 sujets en janvier 2014) pour un total d'environ 9 heures de signaux. Le même matériel a été utilisé pour les deux sessions même si chaque session a été enregistrée dans un endroit différent. Nous avons également utilisé deux questionnaires pour chaque sujet.

Cette étude a été menée avec des gens qui ne sont pas sous tutelle. Ils ont tous accepté de participer à cette recherche et signé une autorisation pour utiliser et reproduire les images et les voix collectées dans le cadre du projet. Pour rencontrer les expérimentateurs, chaque personne a été accueillie individuellement dans une chambre dans la maison de retraite et a été mise au courant de la possibilité d'arrêter l'expérience à tout moment. Certains sujets sont même venus à deux, l'un pour interagir, l'autre pour observer.

3.6 Expérimentateur

Chaque session est enregistrée et réalisée en présence d'au moins trois expérimentateurs du LIMSI. Le premier expérimentateur gère le système, il enregistre aussi les données audio. C'est lui qui supervise le bon déroulement des scénarios et peut envoyer directement des commandes à distance au robot. Le deuxième expérimentateur s'occupe des caméras et de la synchronisation entre l'enregistrement audio et vidéo, c'est le rôle que j'ai joué durant les enregistrements. Le troisième expérimentateur accompagne le participant tout au long de la séance en s'asseyant à la droite du sujet à une distance d'environ 1 mètre. Sa tâche est de répondre aux questions possibles du sujet et de répéter les paroles de NAO que le sujet ne comprend pas. Il intervient lors de l'enchainement des différents scénarios, puis pose aux participants quelques questions sur l'expérience en fin de session. Bien qu'il ne participe pas activement à la conversation entre le sujet et NAO, il faut noter qu'il s'agit une interaction entre 3 parties, donc sujet-robot-expérimentateur. Un exemple de l'interaction enregistrée dans le corpus est montré au-dessous :

1. NAO introduit un nouveau thème de conversation: « Tu sais ? Je suis très fort en jeu de société. On joue à quoi ici ? »
2. Le sujet n'arrive pas à comprendre la question et montre son hésitation en expression faciale et orale: « Hein ? »
3. L'expérimentateur intervient pour répéter la phrase du robot NAO : « Il vous demande à quoi on joue, ici ? »

4. Le sujet tourne vers l'expérimentateur pour l'écouter puis lui demande la confirmation de la phrase qu'il a entendu: « On joue à quoi ? »
5. L'expérimentateur confirme la phrase et l'explique en la reformulant : « Oui. Quels sont les jeux que vous faites ici ? »
6. Le sujet comprend la phrase et se retourne vers le robot pour répondre à sa question : « Ici, je joue … ».

Une ergothérapeute de l'association Approche était également présente lors de tous les enregistrements faisant le lien avec les personnes âgées et les personnels des EHPADs. L'ensemble de positionnement des intervenants est illustré dans la Figure 10.

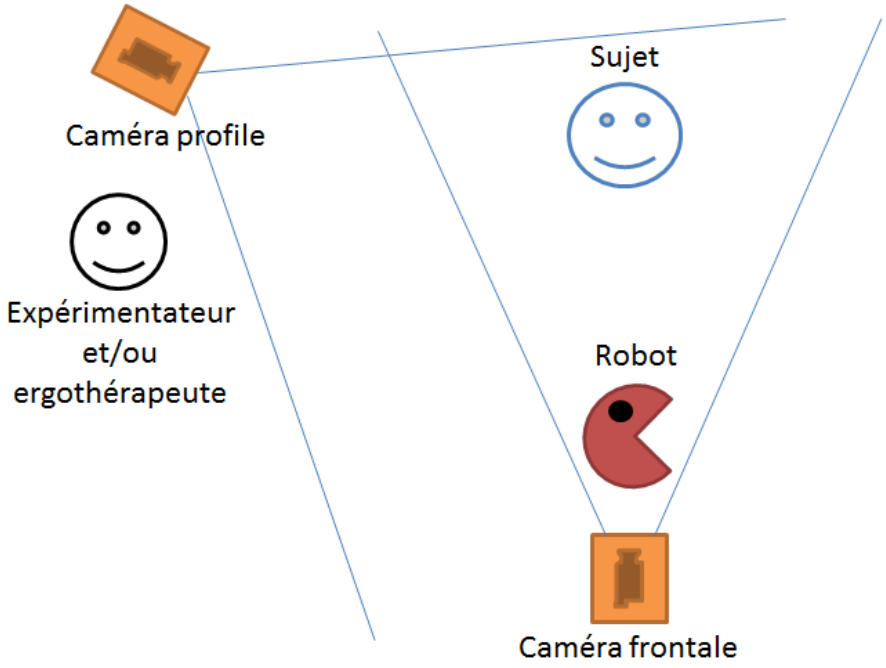

Figure 10 : Positionnement du matériel et des intervenants.

3.7 Déroulement

Chaque session dure au maximum 30 mn. Le robot est assis sur une table en face du participant. Les participants sont invités à discuter avec le robot. Toutes les sessions suivent le même protocole présenté ci-dessous qui se déroule selon cinq étapes.

Etape 0 : Préparation

En préambule au recueil de données, nous présentons au participant l'objectif de l'étude, l'équipe et le cadre de l'étude, et nous lui rappelons qu'il a la possibilité à tout moment d'arrêter l'enregistrement. Un accord de participation, de diffusion et de confidentialité de données dans le cadre de la recherche sur le projet ROMEO est signé. Les données de personnes sont stockées de façon à garder l'anonymat des

participants. Les participants gardent un document signé avec les contacts du groupe de recherche en cas de questions.

Etape 1 : SCENARIO Introduction

Nous commençons par le scénario « introduction » au cours duquel le robot NAO se présente au participant dans l'intention de faire connaissance et lui pose quelques questions, par exemple sur son prénom, comment il souhaite appeler le robot.

Ce scénario vise à familiariser le participant au robot NAO et à sa façon de communiquer. Suite à cette introduction, nous enchainons sur le scénario « Rappel d'évènements »

Etape 2 : SCENARIO Rappel d'évènements

Dans ce scénario, l'agent stimule la mémoire du sujet en l'invitant à se remémorer ce qu'il doit faire à midi. Il ne s'agit pas de donner la réponse au participant immédiatement quand il ne sait pas mais de lui donner des indices afin qu'il arrive par lui-même à trouver la réponse. La suite du scénario est construite de la même manière. L'agent invite la personne à raconter sa vie au sein de l'institution et à partir d'une question très générale « te souviens-tu de ce que tu as fait aujourd'hui ? », il entame une discussion avec le participant et l'aide à se souvenir en faisant des propositions d'activités qui ont pu se dérouler au sein de l'institution.

Ce scénario vise à stimuler la mémoire à court-terme du sujet en reconstruisant les évènements qui se sont déroulés et en lui proposant des indices pour aider le rappel des médicaments.

Etape 3 : SCENARIO Appel d'un être proche

Au cours de ce scénario, le participant est invité par le robot à appeler un de ses proches. Le scénario utilisé vise à interagir avec le participant pour discuter avec lui de souvenirs sur sa famille et spécifiquement sur une personne qu'il pourrait appeler. Le robot devra être en empathie avec le participant. Le robot devra écouter et proposer des réactions positives.

Ce scenario vise l'interaction sociale avec la personne.

Etape 4 : SCENARIO Divertissement et interaction sociale

Au cours de ce scénario, le robot propose de regarder la télévision. Il s'agit d'inviter la personne à discuter sur les programmes télévisés, ce qui lui permet de communiquer son point de vue, d'avoir le sentiment de partager ses impressions concernant les programmes TV et cela permet à nouveau de stimuler la personne.

Après la discussion autour de la TV, le robot joue alors une trentaine de secondes de vieilles chansons françaises et demande au sujet s'il a reconnu le titre de la chanson ou le nom de l'interprète. Ce jeu de reconnaissance permet de stimuler la mémoire à long-terme des personnes âgées.

Ce scenario vise l'interaction sociale avec la personne et la stimulation de la mémoire à long-terme de la personne.

Etape 5 : Remplissage d'un questionnaire de satisfaction

A l'issue de ces jeux de rôles pendant lesquels il interagit spontanément, le participant est invité à donner son point de vue sur le robot et sur la situation d'interaction à travers quelques questions. Deux dimensions seront principalement étudiées à partir de ce questionnaire. Il s'agit d'une part de connaitre la perception des sujets concernant le robot avec lequel ils ont interagi et d'autre part d'avoir leur point de vue concernant la qualité de l'interaction qu'ils ont eue avec le robot.

3.8 Questionnaires

Après chaque interaction, deux questionnaires ont également été utilisés : un premier questionnaire de satisfaction destiné à évaluer la qualité de l'interaction avec le robot, puis une version courte du célèbre questionnaire de personnalité « Big-Five » sous la modèle « OCEAN ». En plus de ces 2 questionnaires, nous avons également des informations fournies par la maison de retraite sur l'âge et le niveau de dépendance mesuré par la grille nationale AGGIR (Autonomie Gérontologie Groupes Iso-Ressources) qui permet d'évaluer le degré de perte d'autonomie ou le degré de dépendance physique ou psychique d'une personne âgée dans l'accomplissement de ses actes quotidiens. L'évaluation de l'AGGIR conduit à positionner la personne âgée dans un Groupe Iso-Ressources (GIR, niveau de dépendance) allant de la dépendance la plus lourde (GIR 1) à l'absence de perte d'autonomie (GIR 6).

3.8.1 Questionnaire de personnalité

Une très brève mesure de la personnalité à la base des « Dix-Point Personality Inventory » (TIPI) [52] a été utilisé. Les questions permettent aux sujets d'évoquer leur perception d'eux-mêmes dans une variété de situations. Le sujet reçoit un ensemble de déclarations et répond en indiquant son degré d'accord avec chaque énoncé sur une échelle de 1 à 7 (1 représente un fort désaccord, 7 représente un accord solide, et les autres valeurs représentent des jugements intermédiaires). Pour chaque sujet, nous avons calculé une valeur pour chacune des cinq dimensions (sous la modèle « OCEAN ») qui sont l'Ouverture à l'expérience, la Conscienciosité, l'Extraversion, l'Agréabilité et la stabilité émotionnelle (contraire de Neuroticisme). Le résultat des personnalités pour les 27 sujets sur une échelle de 1 à 7 avec leur âge

et leur GIR (niveau de dépendance) est montré dans le Tableau 4. Nous pouvons observer que les personnes âgées participant à notre expérimentation ont un âge compris entre 76 et 98 ans, elles ont un niveau GIR et des personnalités très variées d'une à d'autres. Les corrélations avec des analyses précises seront présentées dans la sous-section « 4.1 Analyse des questionnaires ».

Tableau 4 : Personnalités des 27 sujets sur une échelle de 1 à 7 mesurées par le questionnaire de personnalité BIG-FIVE avec leur âge et leur GIR (niveau de dépendance, GIR1 pour la dépendance la plus lourde).

Session	Sujet	Age	Sexe	GIR	E	A	C	S	O
1	1	90	F	2	4,5	5	5,5	3,5	4,5
	2	98	F	5	3	3	5,5	4	4
	3	91	F	1	3	5,5	6,5	4	2,5
	4	78	F	1	2,5	6	5,5	4	3
	5	80	H	6	6	4	7	3	5
	6	86	H	6	2	3	6,5	4	4,5
	7	83	F	3	3	3,5	2,5	6	1,5
	8	80	F	1	2,5	4,5	6,5	1,5	5,5
	9	81	H	1	1	5,5	4,5	4,5	3,5
	10	86	F	2	4	6	3,5	4	4
	11	81	F	3	4,5	3,5	4,5	2,5	6
	12	90	F	2	1,5	5	6	2,5	5
	13	91	F	4	2	6	6,5	4	4,5
	14	86	F	3	2,5	3	6,5	2	4
2	1	88	H	4	4	4	5	1,5	4,5
	2	91	F	4	5,5	3,5	6,5	2	4
	3	80	F	4	6	3	7	5	4,5
	4	90	F	3	6	4,5	3	3,5	5
	5	93	F	2	3	4	6,5	5	3
	6	76	F	4	4	4,5	6	4	7
	7	90	F	4	6	3	5	5,5	3
	8	75	F	4	4,5	5	2,5	2	6,5
	9	89	F	3	3,5	4	4	5	3,5
	10	95	F	4	2,5	6,5	6,5	3,5	4,5
	11	84	F	4	2	7	6,5	5,5	4,5
	12	85	F	5	6	4	7	6	7
	13	89	F	3	4,5	3,5	2	3,5	3

3.8.2 Questionnaire de satisfaction

Pour la première partie du questionnaire de satisfaction, des questions fermées ont été utilisées. On a également demandé aux sujets de fournir des réponses en utilisant un

système d'évaluation sur une échelle de 7 valeurs. Les scores moyens des 27 sujets pour ces questions sont entre parenthèses.

- (Q1) Nao vous comprend-il bien? (5,2)
- (Q2) Nao montre-t-il de l'empathie? (6,3)
- (Q3) Nao est-il agréable de vous? (6,2)
- (Q4) Nao est-il poli? (6,4)

Pour les questions ouvertes, nous donnons une liste d'exemples de réponses ci-dessous. Pour faciliter l'analyse, les réponses ont ensuite été codées en valeurs numériques en utilisant différentes stratégies. Par exemple, pour Q6 nous utilisons 1 pour les noms humains et 0 pour les autres noms. Les valeurs numériques sont utilisées pour calculer les corrélations entre les réponses de satisfaction, les réponses de personnalité et les annotations.

- (Q5) Quel serait le meilleur adjectif pour décrire le robot? (correct, comique, gentil, très agréable, surprenant, amical, drôle, doux, agréable)
- (Q6) Quel nom donneriez-vous au robot? Certains des noms proposés (seulement 4 personnes n'étaient pas en mesure de donner un nom): Pierre, Michel, Alfred, rigolo (bande dessinée), Zizou, Toto, Nano, Nicolin, Jo, gentil (Nice), Patachou, un nom d'un extraterrestre, Mikey.
- (Q7) Vous voulez que le Nao vous vouvoie ou tutoie? 55% des sujets préfèrent la forme familière et 45% disent qu'ils n'ont pas de préférence. Aucun ne préfère la forme soutenue.
- (Q8) Accepteriez-vous de refaire le test avec le robot? 81,5% des sujets sont d'accord.
- (Q9) Aimeriez-vous posséder un robot? Seulement 26% des sujets sont d'accord.
- (Q10) Préféreriez-vous un robot qui ressemble à un robot ou un humain? 55% des sujets préfèrent un robot de type humain.
- (Q11) Considériez-vous le robot comme une machine ou comme un ami ou un compagnon (humain)? La réponse était de 52% pour une machine et de 48% pour une ami et / ou un compagnon.

Les détails sur la corrélation intra-questionnaire et inter-questionnaire seront présentés dans la section « 4.1 Analyse des questionnaires ».

3.9 Annotations

Considérant le contenu du corpus, il existe de nombreuses stratégies d'annotation. Chaque stratégie peut concerner des informations de différents niveaux.

Dans ce travail, nous avons suivi deux schémas d'annotation, un pour l'audio et un pour la vidéo. L'idée est de se focaliser sur le niveau d'engagement du sujet pendant la conversation : à qui il parle, qui il regarde et les signes non-verbaux exprimés ainsi que les expressions émotionnelles, négatives, positives associées. Parmi les signes non-verbaux d'engagement, nous avons étudié principalement le rire et le sourire ainsi que la rotation de la tête. Le rire et le sourire font partie des marqueurs affectifs et sociaux (« affect burst »). Les « affect bursts » sont des expressions émotionnelles spontanées de courte durée, composés d'actions faciales, vocales et gestuelles correspondantes très synchronisées.

Les segmentations et les annotations sont appliquées séparément aux flux audio et aux flux vidéo.

- Pour les flux audio, nous nous sommes concentrés sur des signes non verbaux (rire, toux, souffle, expiration, « tchip », « mmh », etc.) ainsi que sur la source et la destination des signes (e.g. le sujet parle à NAO ou à l'expérimentateur) et sur les informations temporelles de la parole (les bornes des segments). Il existe également une annotation sur les émotions des paroles du sujet qui est détaillée dans la partie analyse.

- Pour les flux vidéo, les annotations concernent deux niveaux : les mouvements du corps et les expressions faciales mais aussi la présence d'émotions (positives, négatives, doute, etc.) lorsqu'elles sont perçues. La partie de l'annotation des mouvements du corps couvrent l'orientation de la tête (gauche, droit, haut, bas), le hochement de la tête (horizontal et vertical), le mouvement global du corps (s'approcher au robot), le mouvement des mains (toucher le robot, manipulation d'objet, cacher le visage, bras croisés, etc.). En ce qui concerne les expressions faciales, le mouvement de sourcils (froncement et haussement des sourcils) et le mouvement de la bouche (rire, sourire avec bouche ouverte, sourire avec bouche fermée, bouche étirée, bouche ouverte surprise, etc.) sont annotées.

L'annotation de la vidéo est faite en utilisant l'outil d'annotation Anvil [70] (voir la Figure 11), celle de l'audio utilise le logiciel de sous-titre « subtitle editor » (voir la Figure 12) dont la sortie est reconvertie sous le format d'Anvil. Les annotations ont été faites par une annotatrice, l'ensemble des annotations sont vérifiées au niveau du contenu des segments et ajustées sur les bornes par un ou plusieurs chercheurs de l'équipe.

La durée totale de données enregistrées pour les 27 sujets est d'environ 9 heures. En raison d'une perte de donnée partielle pour les sujets 7 et 12 de la première session,

les annotations sont faites pour les 25 sujets restants au lieu des 27 sujets initiaux. La période expérimentale (éliminant le temps de réponse au questionnaire, de l'interview, etc.) pour ces 25 sujets concerne environ 6,5 heures de données.

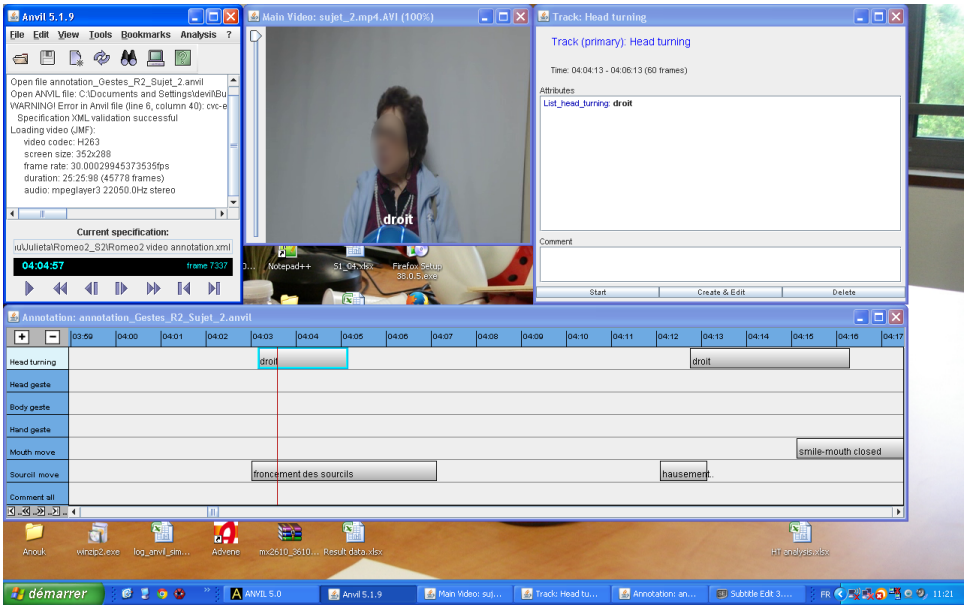

Figure 11 : Sortie d'écran du logiciel d'annotation « Anvil »

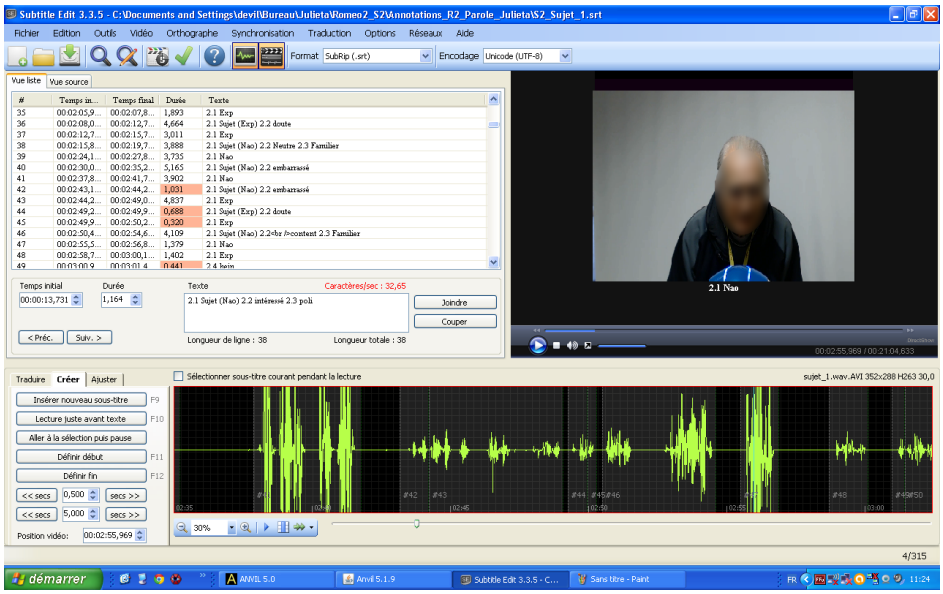

Figure 12 : Sortie d'écran du logiciel d'édition de sous-titre « subtitle editor »

L'annotation audio contient au total 2410 segments de parole du sujet, dont 1881 tours parole des sujets à NAO et 529 tours de parole des sujets à l'expérimentateur. Elle contient également 417 « affect bursts » comprenant 210 évènements de rires.

L'annotation vidéo consiste en 1162 évènements de rotation de la tête avec différentes orientations, 231 évènements de hochement de la tête dont 106 en

horizontal et 125 en vertical, 96 évènements de mouvement du corps, 724 évènement de mouvement de la bouche (y compris 209 rires, 266 sourires avec la bouche ouverte et 98 sourires avec la bouche fermée) et 273 évènements de mouvement de sourcils dont 277 froncements de sourcils et 62 haussement de sourcils. Chaque geste du sujet a été également annoté dans la vidéo avec des étiquettes émotionnelles (positives, négatives, surprise, neutre, etc.).

Les détails plus précis avec des analyses seront présentés dans la sous-section « 4.2 Analyse des annotations ».

4 Analyse

Pour mieux comprendre les corrélations entre les réponses recueillies dans les questionnaires mentionnés dans la section « 3.8 Questionnaires » et le nombre des évènements dans les annotations, des scores sur le coefficient de corrélation linéaire de Bravais-Pearson (« Pearson product-moment correlation coefficient » en anglais) avec un test de permutation sont calculés en utilisant le langage R. Il faut noter que, la valeur de corrélation est accompagnée d'un coefficient de validité (p-value). Une valeur p inférieure à 0,05 signifie un haut niveau de fidélité de la corrélation entre les deux variables. Une valeur p supérieur à 0,05 ne permet pas de conclure de manière décisive.

4.1 Analyse des questionnaires

4.1.1 Corrélations liées à l'âge et au niveau de l'autonomie des personnes

Nous pouvons voir dans la première ligne du Tableau 5 une corrélation positive entre le niveau GIR et la personnalité d'extraversion, qui signifie que les personnes possédant une plus grande autonomie semblent être plus extraverties. La deuxième ligne du Tableau 5 montre que les personnes moins âgées semblent avoir une plus grande ouverture d'esprit, ceci doit être nuancé car le corpus est relativement petit et non balancé sur le genre (27 sujets dont seulement 4 hommes). Les corrélations entre l'âge et les questions 9 et 11 montrent que les personnes moins âgées (vers 75 ans) aiment relativement mieux posséder un robot et elles imaginent plus facilement un robot ayant des caractères humains.

Tableau 5 : Corrélations des réponses aux questionnaires liées à l'âge et au niveau de l'autonomie des personnes. GIR signifie le Groupe Iso-Ressources, allant de la dépendance la plus lourde (GIR 1) à l'absence de perte d'autonomie (GIR 6).

Classe 1	Classe 2	P-valeur	Corrélation
GIR	Extraversion	0,0500	0,381
Age	Ouverture	0,0447	-0,389
	(Q9) Aimeriez-vous de posséder un robot?	0,0195	-0,447
	(Q11) Considériez-vous le robot comme une machine (1) ou un humain (0)?	0,0466	0,386

4.1.2 Corrélations liées à la personnalité

En dehors des corrélations avec l'âge et l'autonomie des personnes âgées, le Tableau 6 montre également les corrélations des réponses aux questionnaires liées à la personnalité. Les deux premières lignes du Tableau 6 montrent qu'il y a une corrélation négative entre l'« extraversion » et l'« agréabilité » pour les personnes

âgées participant à notre expérimentation, et que les personnes âgées plus extraverties perçoivent moins d'empathie du robot Nao. La troisième ligne du tableau montre que les personnes avec une forte amabilité pensent que Nao est plus agréable. Au niveau de la stabilité émotionnelle, selon les lignes 4 et 5, les personnes âgées avec une forte stabilité émotionnelle aiment moins posséder un robot et considèrent le robot comme une machine plus qu'un humain. Les deux dernières lignes du Tableau 6 montrent que les personnes âgées avec un esprit ouvert pensent que Nao les comprend mieux et aiment mieux posséder un robot.

Tableau 6 : Corrélations des réponses aux questionnaires liées à la personnalité. Cor. signifie la corrélation.

Classe 1	Classe 2	P-valeur	Cor.
Extraversion	Agréabilité	0,0231	-0,436
Extraversion	(Q2) Nao montre de l'empathie?	0,0232	-0,435
Amabilité	(Q3) Nao est agréable de vous?	0,0097	0,488
Stabilité émotionnelle	(Q9) Aimeriez-vous de posséder un robot?	0,0004	-0,635
Stabilité émotionnelle	(Q11) Considériez-vous le robot comme une machine (1) ou comme un humain (0)?	0,0221	0,439
Ouverture	(Q1) Nao vous comprends bien?	0,0295	0,419
Ouverture	(Q9) Aimeriez-vous posséder un robot?	0,0197	0,446

4.1.3 Corrélations dans le questionnaire de satisfaction

Dans la première ligne du Tableau 7, on peut trouver une corrélation positive entre le niveau de compréhension du robot et la tendance des personnes âgées à bien vouloir refaire le test. La deuxième ligne montre qu'il est plus facile de percevoir l'empathie pour les personnes âgées quand le robot a une apparence humaine. La troisième ligne prouve que plus le robot est agréable, plus les gens acceptent de refaire le test. En ce qui concerne la quatrième ligne, on peut apprendre que si les gens donnent un nom humain au robot, ils ont tendance à être d'accord sur le fait de posséder un robot. La corrélation de la cinquième ligne signifie que les personnes âgées sont prêtes à tutoyer un robot qui a une apparence humaine, cela signifie également que l'apparence humaine du robot peut réduire la distance mentale (la façon il humanise le robot) entre les personnes âgées avec le robot. Dans les deux dernières lignes du Tableau 7, nous pouvons conclure que les personnes âgées qui considèrent le robot comme un humain ont également plus tendance à bien vouloir refaire l'expérimentation et à posséder un robot.

Toutes ces corrélations prouvent que plus les personnes âgées projettent facilement une image humaine sur un robot, plus elles sont satisfaites de l'expérimentation. De plus, l'apparence humanoïde du robot, la précision de la compréhension du robot et le partage de l'empathie ont un effet positif pour faciliter la projection de l'image d'être humaine sur le robot.

Tableau 7 : Corrélations dans le questionnaire de satisfaction. P. signifie la valeur P. Cor. signifie la corrélation. 1 et 0 sont les valeurs numérisés utilisés pour le calcul de corrélation.

No.	Classe 1	Classe 2	P.	Cor.
1	(Q1) Nao vous comprends bien?	(Q8) Refaire le test avec le robot?	0,0461	0,387
2	(Q2) Nao montre de l'empathie?	(Q10) Préférer une apparence robot (0) ou un humain (1)?	0,0107	0,483
3	(Q3) Nao est agréable de vous?	(Q8) Refaire le test avec le robot?	0,0103	0,485
4	(Q6) Nom humain (1) ou non-humain (0)	(Q9) Posséder un robot?	0,0303	0,417
5	(Q7) vouvoyer (0) ou tutoyer (1)?	(Q10) Préférer une apparence robot (0) ou humain (1)?	0,0124	-0,474
6	(Q8) Refaire le test avec le robot?	(Q11) Considérer le robot comme machine (1) ou humain (0)?	0,0266	-0,426
7	(Q9) Aimeriez-vous de posséder un robot?	(Q11) Considérer le robot comme machine (1) ou humain (0)?	0,0005	-0,627

4.2 Analyse des annotations

4.2.1 Annotation audio

La quantité de conversations annotées dans l'annotation audio du corpus ROMEO2 avec les informations de qui parler à qui est montrée dans le Tableau 8. L'annotation audio contient :

- 1881 tours de parole du sujet au robot correspondant à environ 96 minutes de données et à 24,6% du temps total de l'expérimentation (par rapport à 6,5h) ;
- 529 tours de parole du sujet à l'expérimentateur correspondant à environ 30 minutes de données (7,7% du temps total) ;
- 2029 tours de parole joués par NAO correspondant à 140 minutes de données (35,9% du temps total) ;
- 640 tours de parole correspondant à des réponses de l'expérimentateur au sujet qui ont de questions sur NAO, soit environ 26 minutes de données correspondant à 6,7% du temps total ;
- 109 signaux de bruit soit environ 3,5 minutes de données correspondant à 0,9% du temps total.

De plus, les signaux mineurs comme les voix superposées et la toux du sujet sont annotés. Pour les 2410 tours de parole produits par le sujet, les émotions sont également annotées. Les émotions les plus présentes dans les paroles des sujets sont :

- Neutre : 745 tours de parole (30,9% de la quantité totale de parole des sujets) correspondant à 33,7 minutes de données (26,% de la durée totale) ;

- Joie et satisfaction : 723 tours de parole (30,0% de la quantité totale de parole des sujets) correspondant à 40,7 minutes de données (32,3% de la durée totale) ;
- Doute : 408 tours de parole (16,9% de la quantité totale de parole des sujets) correspondant à 15,6 minutes de données (12,4% de la durée totale) ;
- Colère : 149 tours de parole (7,4% de la quantité totale de parole des sujets) correspondant à 11,2 minutes de données (8,9% de la durée totale) ;
- Surprise : 114 tours de parole (4,0% de la quantité totale de parole des sujets) correspondant à 7,7 minutes de données (6,1% de la durée totale) ;
- Tristesse : 84 tours de parole (3,5% de la quantité totale de parole des sujets) correspondant à 12,3 minutes de données (9,7% de la durée totale).

Tableau 8 : Quantité et durée des conversations annotées dans l'annotation audio du corpus ROMEO2 avec la source et la destinataire de la parole. Exp. signifie l'expérimentateur qui assit à la droite de pour répondre aux questions possibles du sujet et répéter les paroles de NAO que le sujet ne comprenait pas.

Sujet	Source de parole									
	Sujet au robot		Sujet à l'Exp.		Robot		Exp.		Bruit	
Session 1	Nb	Sec	Nb	Sec	Nb	Sec	Nb	Sec	Nb	Sec
1	58	179,2	35	147,0	72	297,9	33	59,4	5	14,6
2	19	49,9	93	614,9	71	291,8	69	177,8	2	1,6
3	53	103,9	19	23,8	79	301,2	25	55,9	2	2,8
4	49	111,0	58	85,6	81	375,6	90	337,8	13	15,5
5	68	195,3	7	27,4	81	369,0	16	34,6	5	8,1
6	40	126,1	27	104,1	64	318,9	22	55,8	1	0,7
8	51	464,3	14	73,4	59	219,7	17	35,2	3	4,6
9	22	25,6	8	6,9	68	355,5	44	152,1	3	19,5
10	68	190,5	5	10,5	88	354,9	10	24,1	5	28,8
11	119	512,3	2	0,9	111	364,9	19	30,1	6	11,0
13	60	206,0	3	12,3	64	336,3	6	16,5	6	13,9
14	64	356,7	1	1,2	61	330,7	12	35,2	3	10,2
Session 2										
1	125	432,9	18	29,9	97	334,0	42	92,8	8	12,1
2	61	169,5	31	51,0	71	355,9	31	66,5	9	10,9
3	126	372,6	10	30,7	89	333,2	14	19,1	5	7,8
4	132	396,2	10	14,4	103	373,7	28	47,5	6	8,2
5	33	52,0	20	25,5	75	398,4	32	78,0	8	11,7
6	94	241,7	4	7,8	79	333,8	3	5,0	5	7,3
7	62	131,3	47	314,1	79	330,9	10	16,1	3	5,1
8	143	428,1	28	64,7	99	273,8	21	46,8	1	0,4
9	94	199,7	29	61,8	101	326,6	39	67,1	3	1,9
10	104	224,4	21	37,9	105	384,3	27	42,3	1	0,6
11	83	133,1	37	64,1	85	366,5	26	47,4	3	3,5
12	57	237,9	1	2,6	59	367,7	3	5,7	2	2,9

13	96	198,6	1	1,0	88	356,1	1	4,8	1	2,4
Total	1881	5739	529	1813	2029	8451	640	1553	109	206
		1:35:39		0:30:13		2:20:51		0:25:53		0:03:26

Nous pouvons observer également dans le Tableau 8 que, le nombre et la durée de parole du robot ne changent pas beaucoup suivant la session en raison du contrôle de la durée de l'expérimentation. En revanche, le nombre et la durée de parole du sujet au robot varient beaucoup selon sa personnalité, la satisfaction liée à l'expérimentation, la compréhension et la stratégie de conversation des différents sujets. De plus, bien que l'expérimentateur qui est assis à la droite du sujet ne participe pas directement à la conversation entre le sujet et le robot, son rôle est de répondre aux questions possibles du sujet et de répéter les paroles de NAO que le sujet ne comprendrait pas, mais certains sujets aiment parler à l'expérimentateur pour raconter des histoires et partager des émotions, dans ce cas-là, l'expérimentateur est obligé de faire une réponse. Le nombre de parole du sujet à l'expérimentateur et le nombre de parole de l'expérimentateur sont donc très variables selon le niveau de compréhension du sujet au robot, la façon il humanise le robot, sa personnalité, sa stratégie de communication et d'autres facteurs.

Tableau 9 : Corrélations entre le nombre de parole du sujet à l'expérimentateur et les autres évènements dans les annotations audio et vidéo.

Classes corrélés	P-valeur	Cor.
Durée sujet parle à l'Exp.	4,92E-08	0,865
Nb parole de l'Exp.	6,17E-05	0,725
Durée parole de l'Exp.	0,001	0,622
Nb de rotation de la tête vers l'Exp. (Anno vidéo)	0,030	0,419
Durée de rotation de la tête vers l'Exp. (Anno vidéo)	0,020	0,446

Les corrélations entre le nombre de parole du sujet à l'expérimentateur et les autres évènements dans les annotations audio et vidéo sont étudiées dans le Tableau 9. Le nombre et la durée de parole du sujet à l'expérimentateur sont évidemment corrélés (première ligne). La deuxième et troisième ligne du tableau montrent que le nombre et la durée de parole de l'expérimentateur sont très corrélés avec le nombre de parole du sujet à l'expérimentateur, cela correspond au fait que l'expérimentateur ne participe pas à la conversation sauf si le sujet le lui a demandé. En ce qui concerne les deux dernières lignes du tableau, les corrélations entre la parole et la rotation de la tête montrent la synchronisation entre le destinataire de la parole (signe auditif) et la direction de regard (signe gestuel), prouvent la possibilité d'utiliser la rotation de la tête comme un indice supplémentaire pour la détection de destinataire de la parole.

L'annotation audio contient également 417 évènements d'« affect bursts », le catalogue des évènements vocaux les plus présents est montré dans le Tableau 10. Le plus présent des « affect bursts » est le rire, qui compte 215 évènements ce qui

correspond à 51,6% de la portion du nombre des « affect bursts » en total. Quand nous comparons le nombre de rire dans l'annotation audio de chaque sujet aux rires annotés dans l'annotation vidéo (décrit dans la partie des annotations de vidéo) dans le Tableau 11, nous avons évidemment une corrélation forte (la valeur P à 7,517e-06 et la valeur de corrélation à 0,778), mais pas 100% de corrélation.

Tableau 10 : Nombre des « affect bursts » vocaux les plus présents dans l'annotation audio.

Sujet	Affect Bursts				
Session 1	Rire	"Ahh"	"Hein"	"Euh"	"Mmh"
1	1	1	3	1	0
2	7	0	2	3	2
3	1	4	3	1	0
4	15	12	0	2	0
5	11	0	0	3	0
6	11	3	0	1	2
8	22	4	0	1	5
9	10	0	0	0	0
10	10	1	0	0	0
11	18	11	8	1	4
13	7	0	0	0	2
14	0	2	0	10	0
Session 2					
1	1	5	10	5	0
2	8	2	0	1	1
3	6	6	1	10	0
4	6	1	0	1	1
5	0	2	0	0	0
6	1	2	0	2	0
7	11	1	3	0	0
8	4	2	0	0	0
9	8	2	13	0	0
10	2	2	0	1	0
11	49	7	0	0	1
12	1	0	0	0	0
13	5	3	0	0	0
Total	215	73	43	43	18

Tableau 11 : Corrélation entre les rires annotés dans l'annotation audio et vidéo.

		P.	Cor.
Nb rire dans l'annotation audio	Nb rire dans l'annotation vidéo	7,52E-06	0,778

4.2.2 Annotation vidéo

Le Tableau 12 montre le nombre des rotations de la tête dans différentes directions et la durée correspondante pour les 25 sujets dans l'annotation vidéo du corpus

ROMEO2. Nous pouvons voir que, comme l'expérimentateur est assis à la droite du sujet, le nombre de rotation vers la droite des sujets dans l'annotation est beaucoup plus important que celui vers les autres directions. La durée où la tête d'un sujet est tournée vers la droite est d'environ 66 minutes donc 16,92% de la durée totale de l'expérimentation (6,5h).

Tableau 12 : Nombre des rotations de la tête et la durée correspondantes pour les 25 sujets dans l'annotation vidéo du corpus ROMEO2.

		Droit		Gauche		Haut		Bas	
Session	**Sujet**	**Nb**	**Sec**	**Nb**	**Sec**	**Nb**	**Sec**	**Nb**	**Sec**
1	1	54	245,40	1	1,80	2	3,27	6	32,73
	2	107	742,76	5	9,47	0	0	10	29,03
	3	29	88,90	1	1,83	0	0	26	171,40
	4	37	258,70	5	12,20	0	0	1	3,23
	6	52	211,76	10	18,23	1	1,27	0	0
	8	27	94,27	2	8,67	0	0	0	0
	9	43	112,93	12	42,73	0	0	22	96,73
	10	19	47,10	1	2,80	0	0	2	6,10
	11	53	168,70	0	0	0	0	0	0
	12	8	20,00	3	3,90	0	0	2	4,67
	13	9	42,03	0	0	0	0	1	3,33
	14	31	104,10	2	5,47	0	0	4	20,17
2	1	27	127,20	0	0	0	0	1	1,77
	2	57	216,96	0	0	0	0	0	0
	3	35	96,87	3	8,53	0	0	3	14,90
	4	54	109,90	3	5,33	0	0	6	31,70
	5	30	131,50	1	2,33	0	0	0	0
	6	12	41,93	0	0	0	0	6	33,43
	7	37	371,70	3	9,17	0	0	15	62,43
	8	74	143,27	6	14,50	0	0	6	21,07
	9	13	89,80	0	0	0	0	1	3,57
	10	35	106,57	0	0	0	0	4	17,67
	11	64	213,23	12	30,40	0	0	25	109,20
	12	1	9,23	0	0	0	0	5	49,17
	13	33	210,46	0	0	0	0	2	5,10
Total		941	4005,26	70	177,36	3	4,53	148	717,39
			1:06:45		0:02:57		0:00:05		0:11:57

Nous avons déjà présenté la corrélation entre le nombre de rotations de la tête vers la droite dans l'annotation vidéo et le nombre de parole du sujet à l'expérimentateur dans l'annotation audio dans la sous-section « 4.2.1 Annotation audio». Nous avons également proposé que la synchronisation entre la destinataire de la parole (signe auditif) et la direction de regard (signe gestuel) puisse permettre d'utiliser la rotation

de la tête comme un indice supplémentaire pour la détection de destinataire de la parole. Dans le Tableau 13, nous montrons l'ensemble des corrélations liées à la rotation de la tête. Toutes ces corrélations démontrent la synchronisation entre l'évènement de rotation de la tête dans l'annotation vidéo avec la personne à qui il parle, soit le robot, soit l'expérimentateur.

Tableau 13 : Corrélations liées à la rotation de la tête dans l'annotation vidéo. Anno. signifie l'annotation. Exp. signifie l'expérimentateur.

Classe 1 (Anno. vidéo)	Classe 2 (Anno. audio)	P.	Cor.
Nb rotation vers le droit	Nb sujet parle à l'Exp.	3,50E-04	0,669
	Durée sujet parle à l'Exp.	8,12E-04	0,637
	Nb parole de l'Exp.	2,51E-02	0,456
	Durée parole de l'Exp.	1,21E-01	0,326
Durée rotation vers le droit	Nb sujet parle à l'Exp.	2,32E-08	0,874
	Durée sujet parle à l'Exp	3,5.5E-11	0,932
	Nb parole de l'Exp	8,10E-03	0,527
	Durée parole de l'Exp.	3,22E-02	0,438

En plus de la rotation de la tête, nous nous focalisons également sur les mouvements de la bouche dans l'annotation vidéo. Comme il s'agit de l'organe le plus expressif dans le visage, l'étude des marqueurs émotionnels liés au mouvement de la bouche est très utile pour la détection des émotions (positive, négative, surprise, etc.). Chaque mouvement de la bouche peut avoir une étiquette émotionnelle :

- 99,2% des évènements de rire et de sourire (bouche fermée et ouverte) sont accompagnés d'une émotion positive ;
- 56,4% des évènements de bouche étirée sont accompagnés d'une émotion doute, et 40% avec une émotion négative ;
- 63,8% des évènements de bouche ouverte surprises sont annotés avec une étiquette émotionnelle de la surprise, et 21,3% avec l'émotion positive.

Le Tableau 14 montre la quantité des mouvements de la bouche dans l'annotation vidéo du corpus ROMEO2. Nous pouvons voir que, parmi les 724 mouvements de la bouche annotés, le rire compte pour 28,87%, le sourire avec une bouche ouverte compte pour 36,74% et le sourire avec une bouche fermée compte pour 13,54%. L'ensemble des mouvements liés au rire et sourire représente 79,14% du total de mouvement de la bouche annotés.

Tableau 14 : Nombre des mouvements de la bouche dans l'annotation vidéo du corpus ROMEO2.

Mouvement de la bouche

Session	Sujet	Rire	Sourire bouche ouverte	Sourire bouche fermée	Bouche étirée	Bouche ouverte surprise	Autres	Total
1	1	3	0	0	2	1	6	12
	2	15	15	3	4	0	3	40
	3	0	4	0	3	4	0	11
	4	0	7	5	0	1	2	15
	6	5	20	3	6	1	2	37
	8	16	19	4	0	0	1	40
	9	7	23	16	2	0	1	49
	10	13	32	0	2	2	0	49
	11	8	17	18	0	10	1	54
	12	1	23	10	9	0	2	45
	13	4	31	0	0	0	1	36
	14	2	2	0	1	0	2	7
2	1	1	0	0	0	0	0	1
	2	4	8	11	1	0	0	24
	3	8	4	3	12	15	0	42
	4	18	9	6	7	3	0	43
	5	0	0	0	0	3	0	3
	6	1	17	1	0	0	2	21
	7	17	10	3	0	4	2	36
	8	8	5	8	1	0	1	23
	9	11	3	0	0	2	0	16
	10	2	1	1	4	6	3	17
	11	40	3	0	1	5	0	49
	12	13	10	4	0	0	0	27
	13	12	3	2	1	2	7	27
Total		209	266	98	56	59	36	724

Afin d'analyser le lien entre le comportement d'un sujet et leur profil, nous calculons la corrélation entre les réponses aux 3 questions du questionnaire de satisfaction et le nombre de sourires et de rires annotés au sein de l'interaction. Comme ce qui est montré dans le Tableau 15, il existe des corrélations positives entre le nombre d'évènement de rire/sourire et l'utilisation de tutoiement, la préférence d'une apparence robot, la considération du robot comme un humain. Nous pouvons conclure que plus un sujet est détendu, plus il exprime de sourires et de rires pendant l'interaction.

Tableau 15 : Corrélation entre la relaxation au niveau de l'expérience pendant l'expérimentation des sujets et leur nombre de sourires et rire annotées lors d'une interaction. 1 et 0 sont les valeurs numérisés utilisés pour le calcul de corrélation.

Nb. évènements	Question	P.	Cor.
Sourire + rire	(Q7) vouvoyer (0) ou tutoyer (1)?	0,044	0,424
Sourire avec bouche ouverte	(Q10) Préférer une apparence robot (0) ou humain (1)?	0,025	-0,465
Sourire + rire	(Q11) Considérer le robot comme machine (1) ou humain (0)?	0,041	-0,429

Malgré la faible quantité d'évènement de bouche étirée, les corrélations correspondantes sont également montrées dans le Tableau 16. A partir des premières quatre lignes du tableau, nous pouvons conclure que, la bouche étirée est souvent utilisée comme un indice d'insatisfaction quand le sujet pense d'être mal traité. De plus, la corrélation négative entre le nombre d'évènement de la bouche étirée avec une émotion doute et la personnalité de l'extraversion est également trouvée.

Tableau 16 : Corrélation entre le nombre des évènements annotés de bouche étirée et les questionnaires de satisfaction et de personnalité.

Nb. évènements	Question et Personnalité	P.	Cor.
Bouche étirée	(Q3) Nao est agréable pour vous?	0,012	-0,496
Bouche étirée avec émotion négative	(Q2) Nao montre de l'empathie?	0,004	-0,560
	(Q3) Nao est agréable pour vous?	0,024	-0,451
	(Q8) Refaire le test avec le robot?	0,016	-0,476
Bouche étirée avec émotion doute	Extraversion	0,018	-0,470

En dehors des évènements de rotation de la tête et de mouvement de bouche, d'autres évènements comme le hochement de la tête, le mouvement du corps, le mouvement de sourcil et les gestes des mains sont également annotés et présentés dans le Tableau 17.

Tableau 17 : Autres évènements à part de rotation de la tête et de mouvement de la bouche dans l'annotation vidéo.

		Hochement de tête		Mouvement du corps	Mouvement de sourcil		Geste des mains
Session	Sujet	Horizontal	Vertical	Approcher le robot	Haussement de sourcil	Froncement de sourcil	
1	1	14	8	0	0	1	7
	2	5	8	4	0	1	19
	3	1	0	1	3	14	4
	4	0	0	3	0	0	8
	6	18	1	1	2	1	19

	8	11	26	30	1	0	9
	9	2	1	2	0	0	11
	10	3	4	0	6	2	5
	11	0	6	2	4	79	2
	12	2	0	0	10	9	2
	13	12	4	0	0	0	5
	14	9	10	10	1	1	23
2	1	1	15	8	1	14	2
	2	2	12	0	5	28	3
	3	2	5	0	13	15	14
	4	4	3	3	1	13	12
	5	0	0	0	0	4	1
	6	0	4	1	1	7	11
	7	2	2	2	1	22	36
	8	0	1	8	0	0	8
	9	2	2	4	0	18	7
	10	1	1	0	5	30	7
	11	10	3	16	2	14	36
	12	4	2	2	1	0	1
	13	1	7	1	5	4	4
Total		106	125	98	62	277	256

Tableau 18 : Autres corrélations que la rotation de la tête et le mouvement de la bouche.

Nb. évènements	Question et Personnalité	P.	Cor.
Hochement vertical de la tête	Stabilité émotionnelle	0,001	-0,607
	(Q11) Considérer le robot comme machine (1) ou humain (0)?	0,039	-0,414
	(Q9) Aimeriez-vous de posséder un robot?	0,021	0,460
Froncement de sourcil avec une émotion négative	(Q7) vouvoyer (0) ou tutoyer (1)?	0,006	-0,532
	(Q8) Refaire le test avec le robot?	0,018	-0,470
	(Q11) Considérer le robot comme machine (1) ou humain (0)?	0,029	0,437
Haussement de sourcil	(Q2) Nao montre de l'empathie?	0,020	-0,462
	(Q3) Nao est agréable de vous?	0,040	-0,412

Les corrélations liées aux évènements autres que la rotation de la tête et le mouvement de la bouche sont montrées dans le Tableau 18. Comme un signe affirmatif, les corrélations liées au hochement vertical de la tête montrent que plus le sujet aime le robot, plus il fait souvent ce geste. De plus, la fréquence de ce geste diminue avec une personnalité de stabilité émotionnelle forte. Au niveau des corrélations liées au froncement de sourcil avec une émotion annotée comme négative, les sujets qui font plus souvent ce geste veulent garder une distance « mentale » (la façon dont il humanise le robot) avec le robot. En ce qui concerne le haussement de sourcil, les

sujets qui font plus souvent ce geste perçoivent moins d'empathie et d'agréabilité du robot.

4.3 Choix des indices de comportement pour le système automatique

L'objectif de ma thèse est l'étude de l'interaction affective et sociale entre des personnes âgées et un robot à partir du corpus ROMEO2, le choix des indices de comportement à étudier dans le système automatique doit considérer plusieurs facteurs : par exemple l'importance des indices pour le déroulement de l'expérimentation, l'apprentissage de l'état mental ou émotionnel du sujet, l'influence sur la stratégie de communication du robot, etc. De plus, le choix des indices doit également considérer la quantité d'évènements annotés dans le corpus, il faut suffisamment de données pour l'entrainement et le test du système. Ma recherche se focalise sur la détection de l'attention et la détection de rire et de sourire, le choix des indices correspondants à ces deux plans seront présentés respectivement dans les paragraphes suivants.

4.3.1 Attention

La détection de l'attention consiste à détecter quand le sujet ne s'adresse pas au robot et à adapter le comportement du robot à la situation. Les outils utilisés en général dans la détection de l'attention sont constitués de deux niveaux : le bas niveau qui utilise la rotation de la tête, le suivi du regard, l'énergie et la qualité de la voix, etc.; et le haut niveau qui utilise l'analyse du contenu de la parole, les expressions faciales, les gestes significatifs, etc. Après avoir considéré les difficultés liées aux personnes âgées et les résultats d'analyse obtenus par l'étude des annotations du corpus, nous nous intéressons à la rotation de la tête au niveau de l'indice visuel. Comme montré dans le Tableau 9, le Tableau 13 et les paragraphes correspondants, il y a une forte synchronisation entre le destinataire de la parole (signe auditif) et la direction de regard (signe gestuel) ce qui peut nous permettre d'utiliser la rotation de la tête comme un indice supplémentaire pour la détection de destinataire de la parole. Au niveau de l'indice auditif, Campbell [21] a montré que la fréquence fondamentale de la voix varie selon la destinataire de la parole. A partir de cette idée, nous avons essayé d'utiliser l'énergie et la qualité de voix pour la détection du destinataire de la parole.

En conclusion, nous utilisons les indices du bas niveau dans notre système de détection de l'attention, donc la détection de la rotation de la tête dans la partie du système visuel et la détection de l'énergie et la qualité de la voix dans la partie du système auditif. Le système final fusionnera ces deux systèmes de mono-modalité pour atteindre une performance de plus haut niveau.

4.3.2 Rire et sourire

La détection de rire et sourire peut être utilisée pour l'étude sur le profil du locuteur et leurs émotions. Nos intérêts se concentrent sur la détection de rire dans la modalité auditive, la détection de rire et sourire dans la modalité visuelle et la fusion des informations des deux modalités afin d'améliorer la performance du système automatique. La forte corrélation entre la relaxation au niveau de l'expérience pendant l'expérimentation des sujets et leur nombre de sourires et rire annotés lors d'une interaction est déjà montrée dans le Tableau 15, nous espérons trouver d'autres corrélations liées à la performance du système automatique après notre expérimentation.

4.4 Difficulté des données de personnes âgées

Après avoir observé les données dans notre corpus des personnes âgées, nous notons qu'il existe des difficultés pour concevoir un système qui marche pour reconnaître les personnes âgées. Les difficultés principales sont classées au-dessous :

- Difficulté pour l'entrainement du modèle statistique : le manque de données de personnes très âgées
- Difficultés pour la détection visuelle : les rides et le relâchement facial, la façon de sourire, le manque d'expressivité
- Difficulté pour la détection auditive : la qualité vocale
- Difficultés pendant l'expérimentation : la variété du temps de réaction, le niveau de compréhension auditive, la perte de la vue, la perte de la contrôle d'émotion pendant l'expérimentation (très triste en parlant de la famille, etc.), les personnalités extrêmes (toujours silence ou ignorant l'expérimentation, etc.)

Toutes ces difficultés rendent encore plus important le défi de la conception du système de reconnaissance affective automatique pour les personnes âgées. De plus, l'évaluation du système doit être également effectuée sur le corpus standard pour comparer avec la performance des méthodes de l'état de l'art.

Synthèse de la Collecte, annotation et analyse du corpus ROMEO2

Pour comprendre efficacement et modéliser le comportement des personnes très âgées en présence d'un robot, des données pertinentes sont nécessaires. Dans cette partie de la thèse, le protocole de collecte de données et les scénarios d'interaction conçus du corpus ROMEO2 ont été présentés. Nous avons décrit également le corpus recueilli (27 sujets, durée totale 9 heures, âge moyen: 85), les annotations et nous avons discuté des résultats obtenus à partir de l'analyse des annotations et de deux questionnaires (un questionnaire de satisfaction, et le Big-Five questionnaire).

La collecte de données est une partie du projet français ROMEO2, le but du projet est de développer un robot humanoïde qui peut agir comme un assistant d'accompagnement pour les personnes souffrant de perte d'autonomie. Notre stratégie était de chercher la population ciblée dans une maison de retraite de personnes âgées, de concevoir plusieurs scénarios de conversation intéressants qui devaient encourager les gens à coopérer avec le robot, et d'utiliser un schéma de Magicien d'Oz (« Wizard of Oz » en anglais) pour contrôler le robot afin de lui permettre d'adapter précisément et rapidement son comportement à la plupart des situations. Après chaque interaction, deux questionnaires ont également été utilisés : un premier questionnaire de satisfaction destiné à évaluer la qualité de l'interaction avec le robot, puis une version courte du célèbre questionnaire de personnalité « Big-Five » (modèle « OCEAN »).

Toutes les annotations ont été menées par une annotatrice, ensuite une vérification des annotations a été menée par plusieurs membres de l'équipe. Au niveau des annotations vidéo, j'ai relu les contenus et vérifié les bornes des segments annotés de notre annotatrice. La correction était faite sur environ la moitié des annotations contenant le rire, le sourire et la rotation de la tête, et majoritairement appliquée (environ 90% des corrections) pour ajuster les bornes des segments annotés. Pour l'annotation du corpus, l'idée est de se focaliser sur le niveau d'engagement du sujet pendant la conversation. A cet effet, nous nous sommes intéressés à la durée de temps pendant laquelle le sujet regarde ou ne regarde pas le robot et au niveau de compréhension du sujet. De plus, de nombreux indices non-verbaux comme le rire, le sourire ont été annotés. Les annotations sont appliquées aux deux flux audio et vidéo. Pour le flux audio, nous nous sommes concentrés sur les signes non verbaux tels que la source, la destination, les informations temporelles, les émotions, et les « affect bursts ». Pour les flux vidéo, les annotations concernent deux niveaux : les gestes du corpus et les gestes faciaux liés aux expressions faciales. Les segments sont également annotés émotionnellement. Au niveau de la durée du corpus annoté, la période expérimentale (éliminant le temps de réponse au questionnaire, de l'interview, etc.) concerne environ 6,5 heures de données.

Pour mieux comprendre les corrélations entre les réponses recueillies dans les questionnaires mentionnés et le nombre des évènements dans les annotations, des scores sur le coefficient de corrélation linéaire de Bravais-Pearson avec un test de permutation sont calculés en utilisant le langage R. Plusieurs phénomènes liés à l'esprit et au comportement des personnes âgées en interaction avec le robot sont trouvées, par exemple :

- Les personnes moins âgées ont tendance à avoir un esprit plus ouvert.
- Plus les personnes âgées projettent facilement l'image d'un humain sur le robot, plus les personnes sont satisfaites de l'expérimentation. De plus, l'apparence humaine, la précision de la compréhension du robot et le partage de l'empathie ont un effet positif pour faciliter la projection de l'image d'un humain vers un robot.
- La durée où la tête d'un sujet est tournée vers la droite est d'environ 66 minutes donc 16,92% de la durée totale de l'expérimentation (6,5h). La synchronisation entre le destinataire de la parole (signe auditif) et la direction du regard (signe facial) peut permettre d'utiliser la rotation de la tête comme un indice supplémentaire pour la détection de destinataire de la parole
- Il existe des corrélations positives entre le nombre d'évènements de rire/sourire et l'utilisation de tutoiement, la préférence d'une apparence robot, la considération du robot comme un humain. Nous pouvons conclure que plus un sujet est détendu, plus il exprime de sourires et de rires pendant l'interaction.
- La bouche étirée est souvent utilisée comme un indice d'insatisfaction quand le sujet pense être mal traité.

L'objectif de ma thèse est l'étude de l'interaction affective et sociale entre des personnes âgées et un robot à partir du corpus ROMEO2, le choix des indices de comportement à étudier dans le système automatique doit considérer plusieurs facteurs. Ma recherche se focalise sur la détection de l'attention et la détection de rire et de sourire. La détection de l'attention consiste à percevoir quand le sujet ne s'adresse pas au robot et à adapter le comportement du robot à la situation. Après avoir considéré les difficultés liées aux personnes âgées et les résultats d'analyse obtenus par l'étude des annotations du corpus, nous nous intéressons à la rotation de la tête au niveau de l'indice visuel, l'énergie et la qualité de voix pour la détection du destinataire de la parole. La détection de rire et sourire peut être utilisée pour l'étude sur le profil du locuteur et leurs émotions. Nos intérêts se concentrent sur la détection de rire dans la modalité auditive, la détection de rire et sourire dans la modalité visuelle et la fusion des informations des deux modalités afin d'améliorer la performance du système automatique. La forte corrélation entre la relaxation au niveau de l'expérience pendant l'expérimentation des sujets et leur nombre de sourires et rire annotées lors d'une interaction est déjà montrée dans le Tableau 15,

nous espérons trouver d'autres corrélations liées à la performance du système automatique après notre expérimentation.

A la fin de cette partie, les difficultés pour la reconnaissance affective automatique des personnes âgées sont analysées. Toutes ces difficultés exigent, hors le corpus ROMEO2, une évaluation du système sur un corpus standard pour comparer avec la performance des méthodes de l'état de l'art.

Partie III : Systèmes automatiques

Dans cette partie, les systèmes de reconnaissance automatique pour les personnes âgées sont individuellement présentés.

Le chapitre 5 « Détection de l'orientation de la tête » présente une méthode conçue à la base de plusieurs détecteurs de visage de différente orientation pour la détection de rotation de la tête. Le système est évalué sur une partie du corpus ROMEO2 et le corpus standard Pointing04. De plus, la dimension sociale de la rotation de la tête est également étudiée dans ce chapitre.

Le chapitre 6 « Détection multimodale de l'attention » utilise la méthode de détection de rotation de la tête présentée dans le chapitre précédant en combinant avec une méthode de modalité auditive pour la détection de l'attention. Le système est testé sur une partie du corpus ROMEO2.

Le chapitre 7 « Détection de rire et de sourire » présente un système audio-visuel pour la détection de rire et sourire. L'ensemble du système est évalué sur une partie du corpus ROMEO2, la détection visuelle est également évaluée sur le corpus standard GENKI-4K. De plus, les corrélations statistiques entre la performance des systèmes et les questionnaires et les annotations sont analysées.

5 Détection de l'orientation de la tête

L'interaction sociale humain-robot a souvent lieu en présence d'autres humains, même si la conversation est censée se dérouler entre le robot et un sujet humain unique. Que ce soit pour s'adresser aux autres locuteurs ou pour toutes autres raisons, le sujet en interaction avec le robot pourrait temporellement détacher leur attention de la conversation. Par conséquent, un système efficace de conversation homme-robot devrait permettre au robot de gérer de telles situations et de s'y adapter. Pour atteindre cet objectif, il y a beaucoup de solutions telles que la localisation de la voix et le suivi du regard. Ce chapitre de ma thèse utilise une technique basée sur la détection de la rotation de la tête. Un sous-ensemble du corpus ROMEO2 est utilisé pour l'évaluation. En définissant quelques règles sur les sorties de deux détecteurs de visage à base de paramètres Haar, mon système a obtenu un score de 81,5% comme résultat de F-mesure et 9,9% comme ratio d'erreur balancé pour la rotation de la tête. Pour connaître les raisons de cette rotation lors de l'interaction orale, le lien entre la rotation et les dimensions sociales ultérieures est analysé. En effet, 68,5% des rotations de la tête ont pour objectif de s'adresser à l'expérimentateur avec des différents motifs (par exemple, demander de l'aide, partager une émotion ou raconter une histoire). Cette expérience suggère que dans les applications de la vie réelle, les données multimodales devraient être utilisées pour mieux comprendre l'interaction orale homme-robot.

5.1 Introduction

Le but de ce chapitre de ma thèse est d'étudier l'interaction sociale entre un robot et les personnes âgées afin de construire un système multimodal qui permette au robot de comprendre le comportement de l'interlocuteur humain et d'adapter ses stratégies d'interaction en conséquence. Ce travail se déroule dans le cadre du projet ROMEO2 (http://projetromeo.com), dont le but est de développer un robot humanoïde qui peut agir comme un assistant complet pour les personnes souffrant de perte d'autonomie. Dans le projet ROMEO, le précurseur de ROMEO2, notre équipe avait déjà développé un système qui identifie et détecte les émotions exprimées par les locuteurs à partir de l'audio [126]. Un profil émotionnel et interactionnel de l'utilisateur [37] est également construit lors de l'interaction et sert de base pour la sélection du comportement le plus pertinent du robot vers l'utilisateur en fonction du contexte d'interaction. Dans les applications de la vie réelle, une interaction homme-robot aura très probablement lieu en présence de plusieurs personnes. Par conséquent, le robot doit détecter si un humain s'adresse à lui. Une solution peut être l'utilisation d'une technique de localisation de voix [80] qui repose sur l'utilisation d'un matériel spécifique. Le suivi du regard peut aussi être utilisé à cet objectif. Par contre, le suivi de regard semble être difficile quand il s'applique aux personnes âgées en raison des rides du visage et de leur tendance à ne pas regarder le robot en raison de problèmes

de vue. En combinaison avec la détection de la parole, la détection de la rotation de la tête peut donc être une alternative intéressante qui permet au robot de se rendre compte du fait que le sujet ne s'adresse en fait pas à lui.

Ce document présente une méthode qui utilise uniquement des détecteurs de visage disponibles dans la librairie OpenCV (http://opencv.org/) pour détecter la rotation de la tête des personnes âgées. Trois détecteurs ont été testés: le détecteur de visage frontal de Haar, le détecteur de visage de profil de Haar, et le détecteur de visage de profil de LBP. Notre approche consiste en une fusion à base de règles qui utilise la sortie des 3 détecteurs de visage comme entrées. Ce système de détection est évalué sur le corpus standard Pointing04 pour comparer la performance avec les méthodes de l'état de l'art et sur notre corpus ROMEO2 de personnes âgées. La collecte du corpus ROMEO2 a été effectuée en présence de quatre expérimentateurs. Un des expérimentateurs était assis à la droite du sujet pour expliquer le processus de l'enregistrement. Après un examen de l'ensemble du corpus, nous avons constaté que la plupart des sujets ont tourné fréquemment leur tête vers l'expérimentateur pendant les séances d'enregistrement. L'idée de base de ce travail est que la détection de la rotation de la tête peut efficacement aider le robot à réaliser si un sujet participe activement à la conversation. Ce chapitre de ma thèse présente une méthode pour la détection de rotation de la tête testée sur le corpus ROMEO2 donc un corpus réaliste d'interaction entre les âgées personnes et le robot.

5.2 Corpus de test utilisé

5.2.1 Sous-ensemble du Corpus ROMEO2

L'ensemble du corpus ROMEO2 contient environ 9 heures de données (une moyenne de 20 minutes par sujet). Dans la première expérience, seulement un sous-ensemble du corpus a été utilisé en raison de l'absence d'annotation complète du corpus. La sous-partie du corpus utilisée contient des segments vidéo de la caméra frontale durant de 60 jusqu'à 120 secondes. Parmi les 27 sujets, 3 sujets ont été écartés de notre expérience pour des raisons différentes (la vidéo frontale d'un sujet a été perdue, un sujet handicapé visuel n'a pas tourné la tête, et un sujet avait la tête tournée de 30 degrés pendant toute l'interaction). Pour les 24 sujets restants, les segments et représentant pour chacun le plus grand nombre d'occurrence de tête tournée ont été sélectionnés et extraits. La durée totale du sous-ensemble est d'environ 30 minutes, parmi lesquelles la durée de rotation de la tête représente 23,8% de la durée totale.

5.2.2 Annotation

Afin d'évaluer le système de détection de l'attention, le sous-ensemble du corpus

d'interaction est annoté. Les segments vidéo frontaux de 1 jusqu'à 2 minutes par sujet ont été présélectionnés afin qu'ils comprennent des paroles du sujet au robot et des paroles du sujet à l'expérimentateur. L'annotation couvre les informations suivantes :

1. Les informations temporelles de toutes les paroles du sujet
2. A qui le sujet parle (robot ou expérimentateur)
3. Les informations temporelles des rotations de la tête
4. La raison de la rotation de la tête :
    - Besoin d'aide ou de confirmation
        a) Le sujet a besoin de l'expérimentateur pour répéter ou expliquer la dernière parole du robot
        b) Le sujet n'est pas sûr de ce qu'il a entendu ou vu
        c) Le sujet essaie de demander des informations à l'expérimentateur, par exemple : la météo, le repas du midi, etc.
    - Faire un contact des yeux
    - Raconter une histoire à l'expérimentateur
    - Partager une émotion

L'effet de la rotation de la tête est un marqueur important pour l'algorithme autonome de NAO. La détection de la rotation de la tête peut aider le robot à distinguer si un sujet est en train de se focaliser sur la conversation, et choisir l'étape prochaine d'interaction parmi l'attente, la répétition de parole, le prochain énoncé ou le fait d'effectuer une interaction émotionnelle. Au niveau des interactions émotionnelles, les gens tournent généralement leur tête en l'accompagnant d'une expression émotionnelle. La reconnaissance de leur état émotionnel actuel sera certainement d'un grand intérêt. Dans une interaction multimodale, la rotation de la tête peut être un indicateur pour passer d'une modalité à l'autre (vidéo, audio ou les deux). Puisque les sujets tournent souvent leur tête vers l'expérimentateur qui est toujours assis sur leur droite dans notre protocole d'enregistrement, un système de la détection de la rotation de la tête à droite est nécessaire.

5.2.3 Corpus standard Poiting04

En complément de l'analyse sur le corpus ROMEO2 des personnes âgées, une évaluation sur un corpus standard est nécessaire pour pouvoir comparer avec les performances de l'état de l'art. Le corpus Pointing04 a été collecté par Gourier et al. [53] au centre de recherche de Grenoble de l'INRIA, où 15 personnes ont reçu l'instruction de regarder successivement 93 marqueurs qui couvrent une demi-sphère en face de la personne.

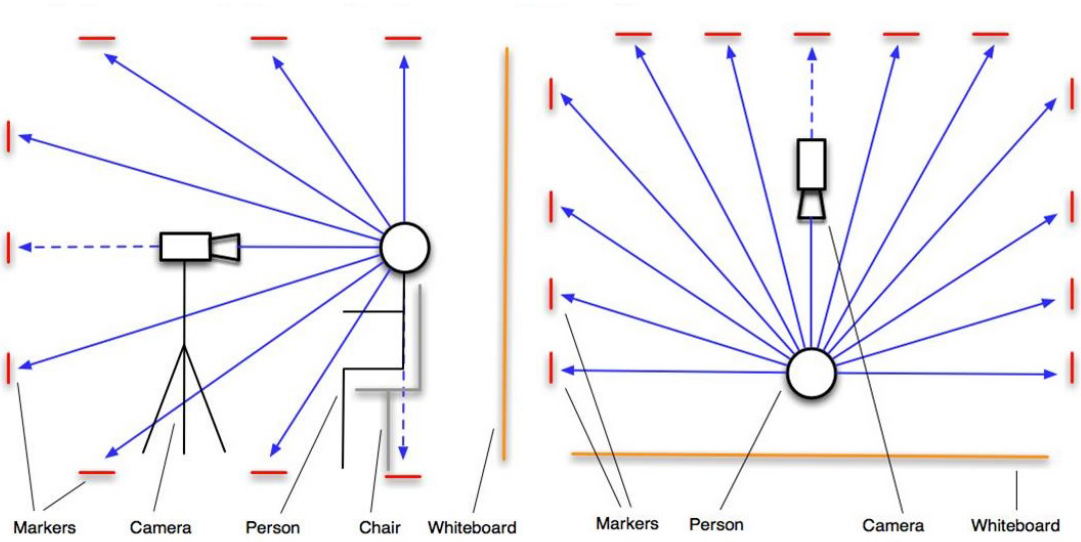

Figure 13 : Positionnement de marqueur en vertical (image à gauche) et en horizontal (image à droite) pour la collection des images des orientations de la tête dans le corpus Pointing04. Les images sont récupérées du site du corpus : http://www-prima.inrialpes.fr/Pointing04/data-face.html

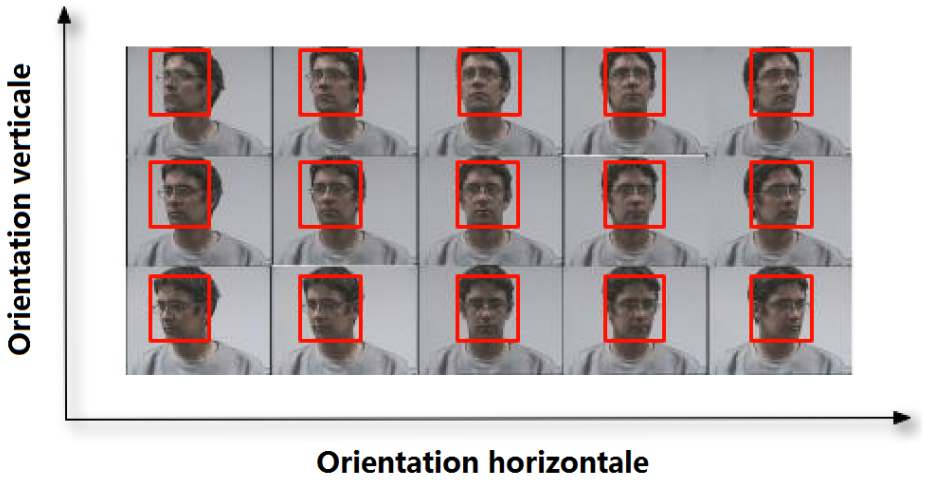

Figure 14 : Exemple des images collectées dans le corpus Pointing04. Les images sont récupérées du site du corpus : http://www-prima.inrialpes.fr/Pointing04/data-face.html

Le corpus se compose de 15 ensembles d'images, chaque ensemble contient des séries de 93 images de la même personne avec différentes positions. Il y a 15 adultes dans la base de données, portant des lunettes ou non et ayant différentes couleurs de peau. Comme cela est montré dans la Figure 13, la position ou la rotation de la tête est déterminée puis collectée en image par 2 angles (horizontal et vertical), qui varient de -90 degrés à +90 degrés. La Figure 14 montre une partie des images collectées dans le corpus Pointing04. Dans mon expérience, comme l'objectif du système est de détecter l'orientation horizontale de sujet, seules les images avec un angle d'orientation verticale entre -30 degrés et 15 degrés et un angle d'orientation horizontale entre -90 degrés et 90 degrés sont utilisés, résultant en 1950 images des 15 sujets.

5.3 Méthode

Le choix de la technique de détection de la rotation de la tête, de notre point de vue, doit considérer les problèmes suivants:

- Il n'y a pas de grand corpus disponible avec des positions de tête différentes des personnes âgées pour l'entrainement statistique d'un système automatique.
- Les personnes âgées ont souvent des problèmes de rides faciales et de relâchement des tissus faciaux. Cela peut influer sur la précision des méthodes basées sur la détection de points caractéristiques dans le visage tels que les modèles actifs d'apparence (« Active Appearance Models » en anglais) [44], les modèles de forme actifs (« Active shape models » en anglais) [34] et d'autres méthodes. En outre, les systèmes existant d'extraction de points sont tous entrainés et conçus pour des images faciales frontales. Cela peut rendre l'utilisation de l'extraction des points dans le visage pour la détection de la rotation de la tête hors de propos.

En considérant tous ces problèmes, ce chapitre de ma thèse présente une méthode qui appartient à la famille des « Detector Array Methods » parmi les 8 catalogues de méthode de détection mentionnés dans l'article [89]. Les trois détecteurs faciaux de différentes orientations horizontales entrainés et disponibles dans OpenCV sont utilisés pour la classification. En cas de détection multiple, ce qui signifie que plus d'un détecteur facial a trouvé un visage, les sorties des détecteurs de visages activés seront comparés pour prendre la décision finale consistant à savoir si il y a une rotation de la tête.

Trois détecteurs de visage ont été utilisés dans ce travail: détecteur de visage frontal de Haar, détecteur de visage de profil de Haar et détecteur de visage de profil de LBP. Ils sont tous disponibles dans la librairie OpenCV et ont été entrainés avec un grand nombre d'images positives et négatives de la population générale. Le détecteur de visage frontal de Haar est entrainé pour reconnaître des images de visage frontal entre environ +/- 40 degrés. Le détecteur de visage de profil de Haar est entrainé pour reconnaître des images de visage de profil gauche (le sujet tourne la tête à droite) à environ 30-70 degrés. Le détecteur de visage de profil de LBP est entrainé pour reconnaître des visages de profil gauche à environ 70-90 degrés. Les 2 détecteurs de visage de Haar sont essentiels à notre système de détection de la rotation de la tête. Pour une meilleure fiabilité, le détecteur de visage de profil de LBP est utilisé comme un détecteur complémentaire. Ses effets sur la performance de l'ensemble du système seront également étudiés.

La vidéo capturée par la caméra frontale est enregistrée à une résolution de 1280x720 pixels et un taux de 29,97 images par seconde. Pour environ 89,75% des frames, au moins un détecteur de visage de Haar a été activé. En outre, au moins un des trois

détecteurs de visage a été activé sur environ 89,76% des frames. En dehors des visages mal détectés, les visages tournants à gauche à plus de 40 degrés et tournants à droite à plus de 90 degrés ne peuvent pas être détectés

Figure 15 : Visages détectés par les détecteurs de visage de l'OpenCV. Circle rouge : visage trouvé par le détecteur de visage frontal de Haar. Circle vert : visage trouvé par le détecteur de visage de profil de Haar. Circle bleu : visage trouvé par le détecteur de visage de profil de LBP.

Dans une situation idéale, un seul détecteur est activé dans chaque trame. En outre, le détecteur activé est précis sur la position et retourne la taille exacte de la tête. Toutefois, dans la situation réaliste, comme celle de la Figure 15, lorsque les trois détecteurs se déroulent en même temps, deux ou trois détecteurs peuvent être activés. Le système de détection globale devrait alors utiliser les sorties des détecteurs activés afin de prendre une décision. Une façon d'atteindre cet objectif est de définir un ensemble de règles sur les paramètres des visages retournés par les détecteurs activés.

En analysant tous les visages détectés dans la vidéo de la caméra frontale, deux modèles peuvent être trouvés:

- Lorsqu'un sujet tourne la tête (à droite), la taille d'un visage détecté par le détecteur de visage frontal de Haar diminue, tandis que la taille d'un visage détecté par le détecteur de visage de profil de Haar augmente. Le taux entre les tailles des 2 visages détectés est essentiel pour la classification de la rotation de la tête.
- Un visage trouvé par le détecteur de visage frontal de Haar est centré sur la position du nez et un visage trouvé par le détecteur de visage profil de Haar est centré sur la face gauche. Nous avons remarqué qu'il y a un écart entre les centres des 2 visages détectés. Cet écart devient également plus grand quand le sujet tourne le visage vers la droite.

Par l'étude de ces modèles liés aux deux détecteurs de visage de Haar, le système de détection de la rotation de la tête quantifie l'intensité de rotation pour chaque trame de vidéo en utilisant les règles suivantes:

1. Si A_{HP} et non $A_{HF.}$ Intensité=1.
2. Si A_{HP} et $A_{HF.}$
    a. Si $S_{HF}*0,1<GAP<S_{HF}*0,4$ et $S_{HP} > S_{HF.}$ Intensité = min $(2*(S_{HP}-S_{HF})/S_{HF.}$ 1);

   b. Sinon Intensité=0.
3. Si non A_{HP} et $A_{HF.}$ Intensité=0.
4. Si non A_{HP} et non $A_{HF.}$ Intensité=Intensité du frame précédant.

(HF signifie le détecteur de visage frontal de Haar, HP signifie le détecteur de visage de profil de Haar, L signifie le détecteur de visage de profil de LBP. Ax signifie l'activation du détecteur X, Sx signifie la taille du visage retourné par le détecteur X, GAP signifie la différence entre le centre horizontal des positions des deux détecteurs de visage de Haar).

La première règle présente le cas où seulement le détecteur de visage de profil de Haar trouve le visage, l'intensité de rotation atteint donc la valeur maximum 1. La deuxième règle concerne le cas où les deux détecteurs de visage de Haar trouvent un visage. Dans ce cas, si la distance entre les centres des deux visages trouvés sont entre 0,1 et 0,4 fois de la taille du visage frontal trouvé, et si la taille du visage de profil trouvé est plus grand que celle du visage frontal, l'intensité de rotation prendra la valeur minimale entre 2 fois la différence de surface entre les deux visage trouvés et 1; Sinon, l'intensité sera 0. En ce qui concerne la troisième règle, seul le détecteur de visage frontal de Haar trouve un visage, l'intensité de rotation sera 0. La règle 4 est pour le cas où aucun des détecteurs de Haar ne trouve un visage, alors l'intensité aura la valeur du frame précédente. En dehors de la possibilité d'une mauvaise détection, la raison qu'aucun des détecteurs de Haar ne trouve un visage peut être que le sujet tourne la tête à gauche ou trop à droite d'une manière qui dépasse la capacité du détecteur de visage profil.

Pour la détection qui utilise tous les 3 détecteurs de visage, une règle supplémentaire doit être appliquée au préalable. Si cette règle ne convient pas, le système suivra les règles mentionnées précédemment :

0. IF A_{L}, Intensité=1.

Afin de lisser la courbe d'intensité au niveau de frame, trois filtres ont été testés pour comparer les performances: le filtre Pascal de 10 éléments, et 2 filtres de moyenne de 5 et 10 éléments respectivement. Pour notre système de détection « online », seules les informations des images précédentes sont utilisées pour le filtrage. Un filtre de n éléments calcule la nouvelle intensité en utilisant l'intensité des n dernières frames y compris le frame actuel. Le filtre Pascal est une extension du triangle de Pascal. La ligne n du triangle de Pascal est une excellente approximation unidimensionnelle de n point d'un filtre gaussien. Le filtre de moyenne prend juste la moyenne des intensités des n derniers frames.

5.4 Evaluation et résultat sur le corpus ROMEO2

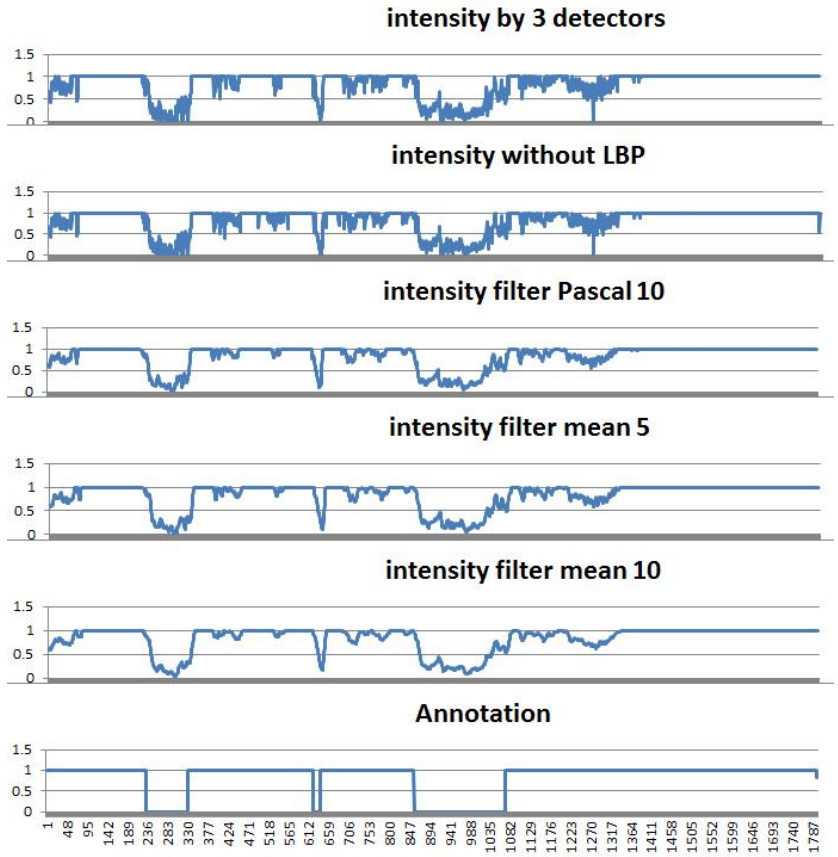

Figure 16 : intensité de rotation de la tête extraite par le système de détection visuelle

La Figure 16 représente les courbes d'intensité de la rotation de la tête extraite au fil du temps pour un sujet dans le corpus ROMEO2. La première courbe est calculée en utilisant les 3 détecteurs de visage, la seconde utilise seulement les deux détecteurs de Haar. La différence entre les deux courbes est très faible. Entre la troisième et la cinquième courbe, les 3 filtres mentionnés dans la sous-section « 5.3 Méthode » sont appliqués sur la première courbe d'intensité pour illustrer les performances du lissage. Nous pouvons constater que tous les filtres fonctionnent bien pour l'élimination du bruit. Les résultats obtenus après une application de différents filtres et l'utilisation de 2 ou 3 détecteurs de visage seront comparées plus tard pour tous les sujets. Pour l'exemple de la Figure 16, 0,5 pourrait être un bon seuil au niveau de l'intensité pour la détection de rotation de la tête. Les détections automatiques sont comparées à l'annotation manuelle pour l'évaluation (voir la sixième courbe dans la Figure 16). Pour les 24 segments vidéo extraits, les occurrences de rotation de la tête vers l'expérimentateur ont été annotés par un logiciel de édition de sous-titres et transformés en étiquettes d'image pour l'évaluation.

L'algorithme mentionné ci-dessus a été utilisé pour chacun des 24 segments vidéo pour trouver le seuil qui conduit à la meilleure performance de détection pour chaque

segment. Le Tableau 19 montre une analyse des meilleures performances de détection au niveau des frames avec les seuils correspondants des 24 segments. Dans ce tableau, les résultats sont obtenus en utilisant les 3 détecteurs de visage, filtre Pascal et le meilleur seuil pour chaque segment. Comme nous pouvons le voir, la moyenne des seuils est de 0,44. La variance du seuil est de 0,13 et elle peut être due à un ou plusieurs facteurs tels que la position de l'expérimentateur, la résolution des images du visage, la variation de la luminosité, l'habitude de rotation de la tête du sujet et de la variabilité de la performance des détecteurs de visage pour les différents sujets. Le ratio moyen de bonne reconnaissance est d'environ 94,2% avec une précision moyenne de 84,2%, un rappel moyen de 88,1%, une F-mesure moyenne de 85,5% et un ratio d'erreur balancé (« Balanced Error Rate » en anglais) de 9,4%. Comme les bords temporels de la rotation de la tête sont difficiles à annoter précisément, la performance de détection peut être légèrement sous-estimée. De plus, comme le nombre de frame de rotation de la tête couvre seulement 23,8% du nombre de frame total, le nombre d'échantillons des classes positive et négative ne sont pas donc balancés, la précision de la détection ainsi que la F-mesure sont également sous-estimées.

Tableau 19 : Analyse des meilleures performances de détection au niveau de frame avec les seuils correspondants pour les 24 segments. BER signifie le ratio d'erreur balancé.

Unité : %	Seuil	Taux de bonne reconnaissance	Précision	Rappel	F-mesure	BER
Max	80,0	94,7	78,5	94,1	85,6	11,4
Min	20,0	95,1	73,7	65,6	69,4	14,7
Moyen	44,4	94,2	84,2	88,1	85,5	9,4
Ecart-type	12,8	3,5	11,5	11,7	9,7	2,6

Le Tableau 20 compare les performances moyennes de détection de l'ensemble des 24 segments vidéo suivant trois axes: le nombre de détecteurs utilisés (deux détecteurs de Haar ou tous les trois détecteurs de visage), le choix du filtre et le type de seuil (global ou par segment). Malgré une performance légèrement meilleure en utilisant les trois détecteurs de visage, la détection avec deux détecteurs de visage de Haar peut aider à réduire la charge de calculs. Ceci est un point important car le système de détection doit être embarqué sur un robot. Les trois filtres résultent tous en un bon lissage, il n'y a que de petites différences de performance. Quant à la F-mesure, les trois filtres ont presque le même résultat avec un léger avantage au filtre de moyenne des 5 dernières trames. Quand le système est appliqué à un sujet inconnu, un seuil commun de 0,5 devrait être envisagé. En utilisant cette valeur de seuil, les performances de détection diminuent d'environ 4% au niveau de la F-mesure et augmentent d'environ 1% au niveau de ratio d'erreur balancé. Le système final, qui utilise les deux détecteurs de visage de Haar et un filtre de moyenne des 10 dernières images et 0,5 comme valeur du seuil, atteint une performance au niveau de la trame de 93,0% en taux de bonne reconnaissance, 85,1% en précision, 81,4% en rappel, 81,5% en F-mesure et 9,9% en

ratio d'erreur balancée. Le score de F-mesure est le même qu'avec la méthode des réseaux de neurones utilisée dans [68] et testée avec des adultes. Par contre, comme ce qui est mentionné, le nombre de trame de rotation de la tête couvre seulement 23,8% du nombre de trames total, la précision de la détection ainsi que la F-mesure du système sont également sous-estimées.

Tableau 20 : Evaluation de la performance de la détection visuelle de rotation de la tête aux trois niveaux. Le système final indépendant du sujet est marqué en couleur bleu.

(Unité : %)	Avec le meilleur seuil pour chaque segment						
	Tous les 3 détecteurs de visage				2 détecteurs de visage de Haar		
Filtre	**PF10**	**MF10**	**MF5**		**PF10**	**MF10**	**MF5**
Taux de bonne Reconnaissance	94,2	94,3	94,3		94,3	94,4	94,4
Précision	84,2	84,8	84,9		85,0	85,4	85,6
Rappel	88,1	87,7	87,8		87,9	87,5	87,6
F-mesure	85,5	85,6	85,7		85,8	85,8	85,9
BER	9,4	9,2	9,1		9,1	8,9	8,8
	Seuil 0,5						
	Tous les 3 détecteurs de visage				2 détecteurs de visage de Haar		
Filtre	**PF10**	**MF10**	**MF5**		**PF10**	**MF10**	**MF5**
Taux de bonne Reconnaissance	92,9	93,0	93,0		92,9	**93,0**	93,0
Précision	83,9	84,7	84,1		84,6	**85,1**	84,8
Rappel	82,4	81,7	82,2		82,0	**81,4**	81,9
F-mesure	81,5	81,5	81,5		81,6	**81,5**	81,6
BER	10,4	10,1	10,3		10,2	**9,9**	10

L'évaluation précédente était au niveau de la trame. Afin d'évaluer le système de détection à un niveau supérieur comme l'évaluation au niveau de segment d'évènement, nous considérons une bonne détection si la majorité des trames dans un segment sont reconnues comme trames de rotation de la tête. Le Tableau 21 montre que 87,7% des segments de rotation de la tête annotés ont été bien détecté. Et peu importe à qui le sujet parle, le taux de détection atteint plus de 83%.

Tableau 21 : Evaluation segmentale de la détection de destinataire adressée

Evènement (fois)	Nombre de rotation de tête annoté	Détection automatique
Parler à l'expérimentateur	77	69 (89,6%)
Parler au robot	36	30 (83,3%)
Regarder ailleurs	1	1 (100%)
Total	114	100 (87,7%)

Dans l'expérience suivante réalisée avec l'annotation complète du corpus, le système est appliqué sur 23 sujets en raison de la perte d'annotation d'un sujet parmi les 24 sujets utilisé pour la première expérience. Le corpus complet utilisé pour l'expérience finale est composé d'environ 6,5 heures de données, la durée de rotation de la tête

représentant 17,3% de la durée totale. La performance au niveau de la trame atteint 92,4% en taux de bonne reconnaissance, 75,3% en précision, 73,0% en rappel, 74,1% en F-mesure et 14,7% pour le ratio d'erreur balancée.

5.5 Evaluation sur le corpus Pointing04

En ce qui concerne l'évaluation sur le corpus Pointing04, seules les images avec la tête en rotation horizontale entre -90 degrés et 90 degrés et en rotation verticale entre -30 degrés et 30 degrés sont utilisées, résultant en 1950 images. A la différence du système testé sur le corpus ROMEO2, le test sur le corpus Pointing04 nécessite la détection de rotation de la tête à gauche et à droite. La méthode de classification est dont légèrement modifiée. Pour la détection de rotation à gauche, comme montré dans la Figure 17, l'image d'origine sera retournée par symétrie d'axe vertical pour obtenir une image reflet. L'application de la détection de rotation de la tête à droite sur l'image reflet peut donc détecter la rotation vers la gauche de la tête. Le système de détection utilise principalement les deux détecteurs de visage de Haar (frontal et de profil) et le seuil 0.5 pour l'intensité de rotation. Comme la source à traiter est constituée d'images isolées, les filtres de lissage ne sont pas appliqués. Si aucun visage n'est trouvé par les détecteurs de visage de Haar, les sorties du détecteur de visage de profil de LBP seront considérées pour déterminer la rotation. Dans le cas où un visage tournant à gauche et un visage tournant à droite sont trouvés dans la même image, la taille des visages de profils seront comparés, et le visage avec la plus grande taille sera dominant pour prendre la décision.

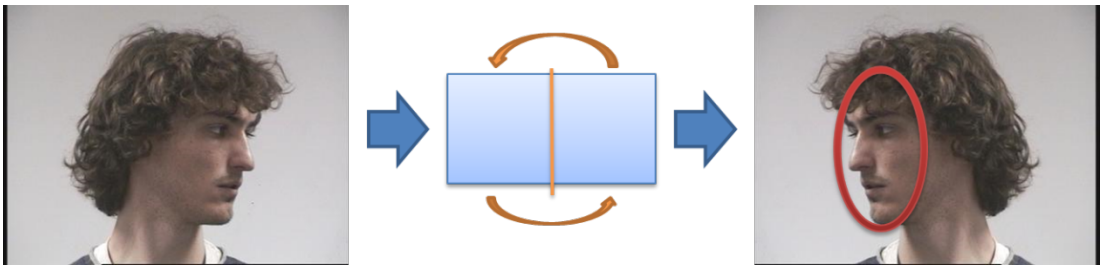

Figure 17 : Détection de rotation de la tête à gauche en utilisant le modèle de détection à droite par le retournement vertical de l'image.

Tableau 22 : Tableau de l'évaluation annotation-prédiction testé du système de la détection de rotation de la tête sur le corpus Pointing04. Le taux de bonne reconnaissance pour chaque classe de rotation horizontale est marqué en couleur bleu.

	Degré annoté												
Prédiction	**-90°**	**-75°**	**-60°**	**-45°**	**-30°**	**-15°**	**0°**	**15°**	**30°**	**45°**	**60°**	**75°**	**90°**
Gauche (%)	**91,3**	**93,3**	**82,7**	**49,3**	13,3	2	0	1.3	0	0	0	0	0.7
Frontal (%)	8	6,7	14	48,7	**86**	**97,3**	**100**	**96**	**72,7**	36,7	18,7	14,7	16,7
Droite (%)	0,7	0	3,3	2	0,7	0,7	0	2,7	27,3	**63,3**	**81,3**	**85,3**	**82,7**

Le résultat de détection du système est montré dans le Tableau 22. Nous considérons que les rotations horizontales entre -90 et -45 degrés correspondent à une tête tournant à gauche, les rotations horizontales entre -30 et +30 degrés correspondent à une tête frontale, et les rotations horizontales entre +45 et 90 degrés correspondent à une tête tournant à droite. Le taux de bonne reconnaissance pour chaque classe de rotation horizontale est marqué en couleur bleu, et sa moyenne est 83,2%. Nous pouvons voir que, la performance de détection est moins fiable pour les têtes avec une rotation de -45 et 45 degrés en raison de la difficulté de déterminer précisément une limite commune entre un visage de profil et frontal pour toutes les personnes.

5.6 Analyse de la dimension sociale de la rotation de la tête

Pour enquêter sur les raisons de rotation de la tête lors de l'interaction orale, le lien entre la rotation de la tête et les dimensions sociales est analysé. Le Tableau 22 montre une répartition de rotation de la tête des sujets en fonction des différents destinataires (ligne) et des raisons (colonne). Parmi les 114 évènements de rotation de la tête dans le sous-ensemble du corpus ROMEO2 utilisé, 68,5% des évènements sont pour parler à l'expérimentateur, dont 38,6% sont pour demander de l'aide ou une confirmation. Ce fait est principalement causé par le problème d'audition des personnes âgées et quelques prononciations inarticulées de NAO. 27,4% des rotations de la tête concernent un contact visuel avec l'expérimentateur ou pour simplement regarder ailleurs. Cela peut être considéré comme faisant partie de la nature humaine, qui est que, peu importe ce qu'ils font, les gens ont toujours tendance à faire attention à l'environnement. 14,9% des rotations de la tête sont pour faire une conversion avec l'expérimentateur et 14,0% sont pour partager l'émotion. Cela signifie que le contenu de nos scénarios de conversation encourage les gens à exprimer leur émotion.

Tableau 23 : Répartition de rotation de la tête des sujets aux différentes destinataires adressées (ligne) et les raisons (colonne).

Raisons (%)	Expérimentateur	Robot	Ailleurs	Total
Besoin d'aide ou confirmation	38,6	0,9	0	39,5
Contacte des yeux avec l'expérimentateur ou regard ailleurs	8,0	18,6	0,9	27,4
Conversation ou raconter une histoire	12,3	2,6	0	14,9
Partage d'émotion	8,8	5,3	0	14,0
Autres	0,9	3,5	0	4,4
Total	68,5	30,9	0,9	100

Dans le Tableau 24, les annotations manuelles sont utilisées pour calculer le ratio moyen de présence de la rotation de la tête et sa durée moyenne dans les quatre événements de parole les plus importants dans une conversation. Par exemple, la première ligne du tableau signifie que, pendant tous les évènements de conversations

du robot au sujet, 14,0% des évènements sont accompagnés avec la rotation de la tête du sujet, et la durée moyenne de rotation de la tête couvre 5,6% de la durée totale de ces évènements. Comme nous pouvons le voir, les gens regardent la plupart du temps le robot tout en l'écoutant (1ère ligne) ou en lui parlant (3ème ligne). Cela pourrait constituer un bon indice pour que le robot considère que, si un sujet ne le regarde pas, il se focalise sans doute sur quelque chose d'autre. Mais il est prématuré de conclure que le fait que le sujet regarde le robot est un signe fort de l'intérêt. En fait, d'après le tableau, un sujet a tendance à tourner la tête avec une probabilité d'environ 60% (environ 50% en tant que la durée moyenne) pendant qu'il est en interaction avec l'expérimentateur. Cela signifie que le reste du temps, le sujet regarde le robot malgré son interaction avec l'expérimentateur.

Tableau 24 : Ratio moyen de présence de la rotation de la tête et sa durée moyenne dans les quatre événements les plus importants dans une conversation

Evènement de conversation (%)	Ratio moyen de présence de la rotation de la tête	Durée moyenne de la rotation de la tête
Robot parle au sujet	14,0	5,6
Expérimentateur parle au sujet	57,7	47,3
Sujet parle au robot	19,4	11,1
Sujet parle à l'Expérimentateur	61,9	54,9

5.7 Conclusion partielle et perspective

Ce chapitre de ma thèse présente une technique de détection de la rotation de la tête sur la base de la définition d'un ensemble de règles sur les sorties de trois détecteurs de visage de différentes orientations. Le système est évalué sur un sous-ensemble du corpus ROMEO2 et a obtenu un résultat de 81,5% comme F-mesure et 9,9% pour le ratio d'erreur balancé au niveau de la trame sur la détection de la rotation de la tête, et un taux de reconnaissance segmentale de l'évènement de rotation de la tête de 87,7%. Le système est également évalué sur le corpus standard des orientations de la tête Pointing04, le taux de bonne reconnaissance pour trois classes (gauche, droit, frontal) est 83,2%.

Pour mieux comprendre les raisons pour lesquelles un sujet tourne la tête pendant la conversation avec un robot, nous essayons d'établir un lien entre la rotation de la tête et de nombreuses dimensions sociales. En effet, 39,5% du temps, un sujet a besoin de l'aide ou de la confirmation de l'expérimentateur. Quelques personnes âgées ont eu quelques difficultés à comprendre le robot, principalement au début de l'interaction. La rotation de la tête est un marqueur important pour l'algorithme autonome de NAO. La détection peut permettre au robot de s'adapter au sujet en choisissant l'action appropriée lorsque l'attention du sujet n'est pas sur la conversation. En outre, comme nous avons constaté, un évènement de rotation de la tête est souvent accompagné d'une expression d'émotion du sujet, et donc, il serait très intéressant que le robot

reconnaisse ces émotions et adapte ses comportements en conséquence. On peut donc conclure que dans une application réaliste d'interaction orale homme-robot, des données multimodales devraient être considérées en raison de la complexité inhérente à ce genre d'interaction.

Ce chapitre de ma thèse montre que la rotation de la tête peut être efficacement utilisée pour détecter la perte de l'attention du sujet dans l'interaction avec le robot. Par contre, quand la tête du sujet fait face au robot, nous ne pouvons pas être sûr que le sujet soit en train de se focaliser sur l'interaction, une détection avec la modalité auditive sera nécessaire pour accomplir la tâche. Dans la section suivante, nous présenterons notre système de détection multimodale de l'attention en utilisant la rotation de la tête ainsi que des indices auditif.

6 Détection multimodale de l'attention

6.1 Introduction

Dans l'interaction orale entre deux personnes, il y a des règles implicites que chaque partie doit suivre afin de maintenir la fluidité de la conversation et mettre en confiance l'interlocuteur. Du point de vue de la personne qui parle, cela implique la capacité à détecter les cas où son interlocuteur n'est pas en train de l'écouter mais, par exemple, d'écouter une tierce personne. Les règles contiennent également la possibilité de reprendre la conversation à partir du point où la conversation a été interrompue, etc.

Bien qu'il soit très difficile d'imaginer un système d'interaction homme-machine qui suive parfaitement l'ensemble de ces règles d'interaction humains-humains, un système couvrant même un très petit sous-ensemble de ces règles est de nos jours une création très souhaitable. La plupart du temps, dans les systèmes d'interaction homme-machine, la partie humaine doit fournir un effort important pour compléter efficacement l'interaction. Dans l'interaction homme-machine, si la machine est un robot humanoïde, l'attente de l'homme sur la performance de la machine peut devenir beaucoup plus élevée qu'actuellement.

Une utilisation typique de robot humanoïde est dans l'interaction sociale avec les personnes âgées. Ce chapitre de la thèse présente une technique audio-visuelle pour détecter l'attention d'un sujet âgé qui interagit avec un robot. Nous utilisons un sous-ensemble du corpus d'interaction sociale ROMEO2. Le corpus consiste en enregistrements de personnes âgées qui interagissent avec le robot humanoïde NAO avec l'aide d'un psychologue (à qui nous nous référons comme l'expérimentateur) et un ingénieur exécutant un Magicien d'Oz. Comme cela est montré dans le chapitre 5 « Détection de l'orientation de la tête », presque tous les sujets avaient tendance à tourner la tête vers l'expérimentateur qui est assis à côté d'eux, soit pour demander de l'aide ou tout simplement pour partager des pensées. En dehors de la rotation de la tête, nous avons également remarqué une différence entre la façon dont un sujet parle au robot et à l'expérimentateur.

L'idée principale du système de détection de l'attention est d'utiliser la détection de la rotation de la tête mentionnée dans la partie précédente et l'évaluation de la qualité de la voix pour reconnaître si le robot est le destinataire de l'intervention de la personne âgée ou non, et pour adapter le comportement du robot à la situation en conséquence. Ce travail a été réalisé en collaboration avec mon collègue Mohamed SEHILI qui s'occupe principalement de la conception et de la réalisation du système de détection auditive. La détection de la rotation de la tête utilise la fusion des règles appliquées sur la sortie des 3 différents détecteurs de visage : un détecteur Haar de visage frontal, un détecteur Haar de visage de profil et un détecteur LBP de visage de profil. Au

niveau de l'évaluation de la qualité de la voix, un ensemble de signaux audio et « Pitch » liés aux indices sont utilisés pour déterminer si le sujet est en train de parler au robot ou à l'expérimentateur.

6.2 Corpus de test utilisé

Le corpus de test utilise le même sous-ensemble du corpus ROMEO2 utilisé pour la détection de rotation de la tête présenté dans la sous-section « 5.2 Corpus de test utilisé ». Les segments sélectionnés ont été extraits des 24 sujets et représentent la plus grande occurrence de tête tournée pour chacun. La durée totale du sous-ensemble est d'environ 30 minutes, la durée de rotation de la tête représentant 23,8% de la durée totale. Les segments vidéo frontaux de une à deux minutes par sujet ont été présélectionnés afin qu'ils comprennent des paroles du sujet au robot et des paroles du sujet à l'expérimentateur. L'annotation couvre les informations temporelles des paroles du sujet, le destinataire adressé de la parole, les informations temporelles des rotations de la tête et la raison de la rotation de la tête.

6.3 Méthode

6.3.1 Méthode de la détection auditive

Dans ce travail, un ensemble d'indices liés à F0 ainsi que d'autres indices perceptifs audio sont utilisés pour détecter l'attention d'un sujet au moyen d'une analyse de la voix, qui vise à faire une distinction entre la parole du sujet au robot et la parole du sujet à l'expérimentateur.

Un ensemble de 8 types d'indices perceptuels audio ainsi que des indices liés au « pitch » sont utilisés avec un classifieur SVM linéaire pour distinguer le destinataire de la parole du sujet entre le robot et l'expérimentateur. Les indices perceptifs sont calculés sur une fenêtre de 20 ms avec 10 ms de décalage. Au niveau des indices liés au « pitch », nous calculons le taux entre des trames de voix et non-voix, le nombre de pause dans la voix, le « jitter » relatif et absolu, le « shimmer » et le « shimmer » en décibels de chaque région de voix. Les indices de « jitter » et « shimmer » caractérisent le « pitch » et ils représentent les variations de F0 du cycle à cycle et la forme d'onde respectivement [48]. Une séquence de trames est ensuite intégrée dans un vecteur d'indice pour calculer l'ensemble de coefficients statistiques. Le Tableau 25 montre les indices acoustiques et statistiques utilisés. Ces indices acoustiques sont extraits en utilisant la librairie d'indice audio yaafe [84] et aubio[13] (uniquement pour le « pitch »).

[13] http://aubio.org/

Tableau 25 : Indices acoustiques utilisés pour la détection de l'attention

Indices acoustique [# indices par frame] (traduction en français)	Indices statistiques appliqués aux indices acoustiques
Loudness [24] (intensité sonore)	Min, Max, Moyen, Median, STD, Slope, Centroid, Spread, Skewness, Kurtosis
Spectral Flatness [1] (planéité spectrale)	
Perceptual Sharpness [1] (finesse/acuité perceptive)	
Perceptual Spread [1] (étalement perceptif)	
Spectral Roll-Off [1] (atténuation spectrale)	
Spectral Decrease [1] (décroissance spectrale)	
Spectral Variation [1] (variation spectrale)	
Zero Crossing Rate [1] (taux de passage par zéro)	
Paramètres globaux liés à la fréquence fondamentale: proportion de trames voisées, ratio voisé/non-voisé, nombre de ruptures vocale, degré des ruptures vocales)	
Paramètres locaux liés à la fréquence fondamentale (calculés par région voisée): moyen, median, min, max, range (intervalle), slope (pente), jitter, shimmer	Moyen, STD

6.3.2 Méthode de la détection visuelle

La détection visuelle utilise le système de la détection de rotation de la tête mentionnée dans la chapitre 5 « Détection de l'orientation de la tête ».

6.3.3 Méthode de la détection multimodale de l'attention

L'utilisation de plusieurs modalités augmente souvent la précision et la fiabilité d'un système de décision. L'un des objectifs du travail de ma thèse est de montrer le potentiel d'un système audio-vidéo de détection de l'attention qui pourrait dépasser un système monomodal. Il y a plusieurs façons de combiner deux ou plusieurs systèmes pour parvenir à une décision. Dans ce travail, nous utilisons un schéma simple de cascade de deux couches de décision pour la reconnaissance du destinataire à qui le sujet parle en vis-à-vis du robot.

Une hypothèse naturelle qui vient à l'esprit est que les sujets tournent très probablement la tête quand ils s'adressent à l'expérimentateur qui est assis à côté d'eux, alors qu'ils regardent la plupart du temps le robot quand ils parlent au robot. Par conséquent, nous utilisons les règles suivantes pour la détection audio-visuelle de l'attention. Pour l'énoncé de chaque sujet:

- Si le nombre de trames de la tête tournant à droite / le nombre total de trames $\geq \theta$, alors nous considérons que le sujet parle l'expérimentateur (ignorer la décision du système audio).

- Sinon, la décision suive le système audio décision audio-basé.

Le seuil θ représente la quantité de temps où un sujet tourne la tête tout en parlant au robot (appliqué à chaque énoncé). Ces deux règles sont basées sur l'hypothèse ci-dessus et n'ont pas besoin d'entrainement statistique. Évidemment, d'autres stratégies de fusion peuvent être considérées comme la logique floue ou un ensemble de règles plus complexes.

6.4 Résultats expérimentaux

6.4.1 Résultat du système auditif

Comme ce qui est mentionné dans la sous-section « 6.2 Corpus de test utilisé », toutes les paroles des sujets dans les segments sélectionnés sont manuellement horodatées. L'intérêt principal de la partie audio de cette expérience n'est en fait pas d'une détection d'activité de discours, mais l'analyse de signaux vocaux. Une validation croisée par tiers (« 3-fold ») indépendante du sujet a été réalisée avec 24 sujets. Chaque sujet a des énoncés qui appartiennent soit à la classe « au robot » ou « à l'expérimentateur ». La stratégie d'apprentissage indépendant de sujet signifie qu'aucune donnée de l'entrainement et du test n'appartiennent au même sujet. Par conséquent, nous faisons l'hypothèse implicite que les données provenant d'un sous-ensemble de sujets pourraient servir de l'entrainement pour évaluer un autre sous-ensemble de sujets totalement différents. Cela a été effectivement motivé par ce que nous avons observé sur les données collectées: la plupart des sujets ont adopté une stratégie comparable pour s'adresser au robot, ils parlaient souvent plus fortement, lentement et clairement par rapport à ce qu'ils ont fait quand ils s'adressent à l'expérimentateur. De plus, ils ont presque imité la façon de parler du robot.

Le Tableau 26 montre les résultats de validation croisée obtenus. Comme nous pouvons le voir, il y a une bonne reconnaissance des paroles adressées au robot par rapport à ceux adressés à l'expérimentateur. Ceci pourrait être expliqué par le fait que la variation inter-sujet est plus grande en parlant à l'expérimentateur qu'au robot.

Tableau 26 : Détection de destinataire de parole par audio

Destinataire de parole	Expérimentateur	Robot	Moyen
Moyen	33,12%	81,92%	57,52%
Ecart-type	1,99%	7,12%	3,47%

6.4.2 Résultat du système visuel

La performance du système de la détection de rotation de la tête a déjà été présentée dans la sous-section « 5.4 Evaluation et résultat ». Le système final, qui utilise les deux détecteurs de visage de Haar et un filtre de moyenne des 10 dernières images et

0,5 comme la valeur du seuil, atteint une performance au niveau de frame de 93,0% comme le taux de bonne reconnaissance, 85,1% comme la précision, 81,4% comme le rappel, 81,5% comme le F-mesure et 9,9% comme le ratio d'erreur balancé. Au niveau de l'évaluation segmentale, 87,7% des segments de rotation de la tête annotées ont été bien détecté. Et peu importe à qui le sujet parle, le taux de détection atteint plus de 83%.

6.4.3 Résultat de la détection multimodale

Dans la dernière expérience, nous avons mené une validation croisée en utilisant les mêmes sous-ensembles que dans la sous-section « 6.3.1 Méthode de la détection auditive ». Nous avons utilisé le système de décision en cascade décrit en « 6.3.3 Méthode de la détection multimodale de l'attention ». Toutefois, afin de mieux évaluer l'amélioration que le système auditif peut apporter, nous utilisons une détection manuelle de la rotation de la tête comme entrée (plutôt que la détection automatique). Le seuil θ est fixé à 0,5. Par conséquent, si un sujet tourne la tête au moins 50% du temps en parlant, le système va supposer qu'il ne s'adresse pas au robot, mais à l'expérimentateur. Dans le cas contraire, le système considère la sortie du système auditif.

Tableau 27 : Détection audio-visuelle de la destinataire adressée

Destinataire Adressée	Expérimentateur	Robot	Moyen
Moyen	81,48%	76,13%	78,80%
Ecart-type	14,20%	9,21%	3,72%

Le Tableau 27 montre les résultats obtenus par le système multimodal. En comparaison avec le Tableau 26, il y a une grande amélioration de la performance, en particulier pour détecter si le sujet s'adresse à l'expérimentateur. La rotation de la tête semble être corrélée avec le fait que le sujet ne s'adresse pas au robot.

6.5 Conclusion partielle et perspective

Ce chapitre de ma thèse présente un système de détection de l'attention dans le contexte de l'interaction sociale entre les personnes âgées et le robot. L'objectif est de permettre au robot de détecter si le sujet est attentif au robot dans l'interaction et d'adapter son comportement au sujet. L'idée de base est d'analyser les habitudes de comportement du sujet et de faire la distinction entre leur comportement quand ils s'adressent au robot et quand ils sont à l'écoute ou s'adressent à quelqu'un d'autre.

Nous avons d'abord mis en place un système auditif qui analyse la voix du sujet et détecte si le sujet est réellement en train de parler au robot. La détection de la rotation de la tête est aussi un moyen très intéressant pour détecter la perte d'attention du sujet. En fait, un sujet peut tourner la tête pour écouter un autre être humain, sans pour

autant parler lui-même, ce qui rend l'utilisation du système d'analyse par la voix seule très difficile. Nous avons ensuite montré le potentiel d'une méthode en cascade qui utilise à la fois les modalités d'une manière complémentaire. Cette méthode donne de meilleurs résultats par rapport au système auditif.

À l'avenir, nous devrions considérer l'utilisation d'un moyen plus efficace pour combiner des signaux sonores et visuels. Par exemple, on peut imaginer l'utilisation de méthodes qui donnent en sortie une probabilité (par exemple un SVM probabiliste) au lieu d'une décision binaire. Nous devrions également annoter et évaluer une plus grande quantité de données et exécuter une évaluation dépendante du sujet pour savoir si elle peut amener une amélioration.

7 Détection de rire et de sourire

7.1 Introduction

Le rire et le sourire sont considérés comme des indices importants dans la communication humaine. Ils véhiculent beaucoup d'informations au cours de l'interaction humain-humain, telles que l'état émotionnel, la stratégie sociale de la communication et de la personnalité. Ce type d'information apparaît également dans l'interaction homme-robot, en particulier avec des robots humanoïdes. Ce chapitre de la thèse se concentre sur la détection du rire et du sourire des personnes âgées qui interagissent avec le robot NAO dans une situation de la vie réelle. A cet effet, nous utilisons une partie de notre corpus d'interaction sociale ROMEO2 [116].

Chacune des approches audio et vidéo pour la détection de sourire et de rire a ses avantages et inconvénients. En fait, l'audio est une modalité adaptée à la détection de rire, surtout quand un sujet n'est pas en face de la caméra même s'il est possible de rater le son des rires de courte durée. La détection visuelle peut cependant être un bon moyen pour la détection de sourire et de rire quand le sujet est face à la caméra et qu'il n'y a pas d'obstacle entre le sujet et la caméra. Un autre problème envisagé par la détection visuelle de sourire est la similarité entre une bouche souriante et une bouche parlante. La recherche de [101] montre que l'ajout de l'information visuelle (l'orientation de la tête et l'expression faciale) peut légèrement améliorer les performances du système audio. L'utilisation conjointe des deux canaux audio et vidéo pour améliorer la performance globale de la détection est étudiée dans ma thèse.

En dehors des problèmes du corpus réaliste, il faut aussi prendre en compte la difficulté de la détection liée à l'âge des sujets. L'ensemble des problèmes entre le manque de données pertinentes réalistes des personnes âgées interagissant naturellement avec un robot et les difficultés liées aux rides du visage et à la qualité de la voix sont les défis supplémentaires abordés dans ce travail. Nous considérons donc l'utilisation de ces données avec les méthodes de détection proposées comme une partie de la contribution principale de la thèse. En effet, cette partie de travail est en collaboration avec mon collègue Mohamed SEHILI, qui s'occupe principalement de la conception et de la réalisation du système de détection auditive.

7.2 Méthodes

7.2.1 Fusion audio-visuelle pour la détection de rire et sourire

L'efficacité de la détection de sourire à partir d'images a été prouvée sur de nombreux corpus contenant des émotions actées corpus [38, 59, 73, 120]. Cependant, cela n'a pas été étudié sur des données réalistes d'interaction entre des personnes âgées et un

robot. En dehors du problème des rides faciales liées de manière inhérente à l'âge des personnes, la distinction entre une bouche souriante et une bouche parlante soulève un autre défi important pour la détection visuelle. Pour surmonter ce problème, la détection de la parole dans l'audio est utilisée dans le système audio-visuel pour aider le système de détection visuelle à se focaliser uniquement sur la détection de sourire alors que le sujet ne parle pas.

7.2.2 Détection visuelle de rire et sourire

Pour le choix de la méthode de détection visuelle, j'ai fait une analyse sur les méthodes de l'état de l'art :

- Selon FERA11 [142], les méthodes utilisant le suivi des points d'intérêt subissent plus d'erreurs de détection et sont moins robustes lors de l'initialisation. Ces problèmes seront fort probablement aggravés avec les données de personnes âgées.
- Un avantage majeur de l'approche à base d'apparence est l'économie en coût de calcul.
- Dans notre corpus des personnes âgées, la luminosité est plutôt uniforme, l'utilisation des filtres orientés comme Gabor ne donne probablement pas d'avantage.
- A cause des différences en apparence et en temps de réaction entre les expressions spontanées et actées/posées, les méthodes utilisées pour la reconnaissance des expressions actées ou posées pourraient ne pas fonctionner correctement sur les expressions réalistes.

Le système de détection visuelle utilise donc la méthode Local Binary Patterns (LBP) [50, 162] pour l'extraction des indices et un classifieur Support Vector Machines (SVM) [35] avec un noyau Radial Basis Function (RBF) pour la classification.

Les données vidéo étaient enregistrées en 30 images par seconde avec une résolution de 1280x720 pixels. Pour chaque trame, une image frontale du visage du sujet a été extraite en utilisant le détecteur facial de Viola et Jones [145].

Après une égalisation d'histogramme au niveau de l'intensité global de niveau de gris, les images faciales ont été ré-échantillonnées à 64x64 pixels et traitées à l'aide de l'analyse LBF avec une grille de 10x10 pour obtenir un vecteur d'indices pour chaque frame. La performance du classifieur SVM du noyau RBF a été optimisée en testant la configuration de la valeur ***C*** et ***gamma*** entre 0,0001 et 1000 sur une grille logarithmique de facteur multiplicatif de 3.

7.2.3 Détection auditive de rire

Comme le flux audio d'une interaction homme-robot est constitué en général de la parole du sujet, de rires, de la parole du robot et de silence, nous avons utilisé un

système de classification à 4 classes. La classification auditive est faite au niveau des trames à l'aide des 13 coefficients de MFCC et avec 4 modèles GMM afin de représenter ces 4 classes. Chaque trame est de 20 ms de longueur avec un décalage de 10 ms par rapport à la trame précédente.

7.3 Corpus de test

7.3.1 Evaluation sur le corpus standard GENKI-4K

En plus des évaluations sur notre corpus réaliste ROMEO2, les méthodes de détection développées dans ce travail ont été également évaluées sur le corpus acté GENKI-4K [138] pour comparer leur performance avec l'état de l'art. Le corpus GENKI-4K contient 4000 images de portait collectées sur l'internet. Le détecteur de visage proposé par Viola et Jones [145] arrive à trouver 3820 image faciale parmi ces 4000 images. Le protocole de validation croisée est appliqué dans l'évaluation de la détection de sourire. Les 3820 images sont divisées dans 10 groupes de 382 images, les images de chaque groupe sont testées par le modèle SVM entrainé par celles des 9 autres groupes. Le taux de bonne reconnaissance sur ces 3820 images est supérieur à 83%.

Figure 18 : Exemples des deux images de sourire (gauche) et des deux images de non-sourire (à droit) dans le corpus GENKI-4K.

7.3.2 Sous-partie du Corpus ROMEO2 utilisée

Au niveau de la sélection de la sous-partie du corpus ROMEO2 à utiliser pour l'expérimentation du système automatique, dans la partie visuelle, seuls les événements évidents comme le rire et sourire avec la bouche ouverte sont étudiés. Cela concerne à environ 575 événements des 27 sujets dans l'annotation vidéo. Etant donné que la répartition des événements n'est pas équilibrée entre tous les sujets, tous les participants de sexe masculin (3 sujets) et tous les sujets possédant moins de 10 événements de rire et de sourire avec la bouche ouverte ont été rejetés.

En conséquence, notre sous-ensemble de données expérimentales dans l'annotation vidéo consiste en 15 sujets féminins avec 168 évènements de rires et 221 évènements de sourire avec la bouche ouverte, donc 389 événements au total. La durée maximale,

minimale et moyenne des évènements est 15,840, 0,633 et 3,837 secondes respectivement. La durée totale de ces 389 évènements de rire et sourire avec la bouche ouverte est de 1492,4 secondes, ce qui correspond à 10,79% du temps total de la sous-partie du corpus ROMEO2 utilisée puisque la durée totale de l'expérimentation pour les 15 sujets est de 13835 secondes. La répartition du nombre d'évènement de rire et de sourire évident est illustrée dans la Figure 19.

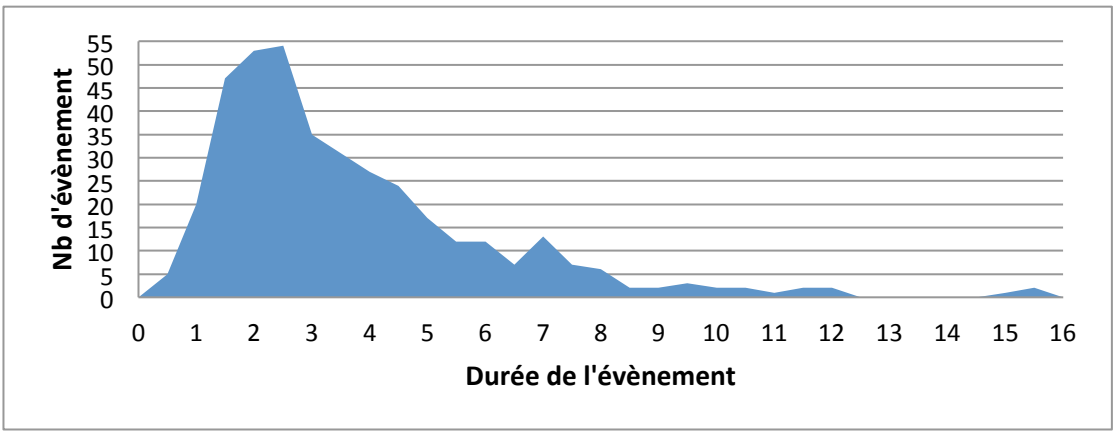

Figure 19 : Répartition de nombre d'évènement de rire et sourire avec bouche ouverte dans l'annotation vidéo de la sous-partie du corpus ROMEO2 des 15 sujets utilisés pour l'expérimentation du système automatique. Le nombre d'évènement est calculé par un intervalle de durée de 0,5 second.

Les annotations audio de ces 15 sujets sont également utilisées pour l'entrainement de modèle statistique du système auditif de la détection de rire. Cette partie de l'annotation audio contient 170 évènements de rire, dont la durée maximale, minimale et moyenne est 6,32, 0,26 et 1,3 second respectivement. La répartition de nombre d'évènement de rire est illustrée dans la Figure 20.

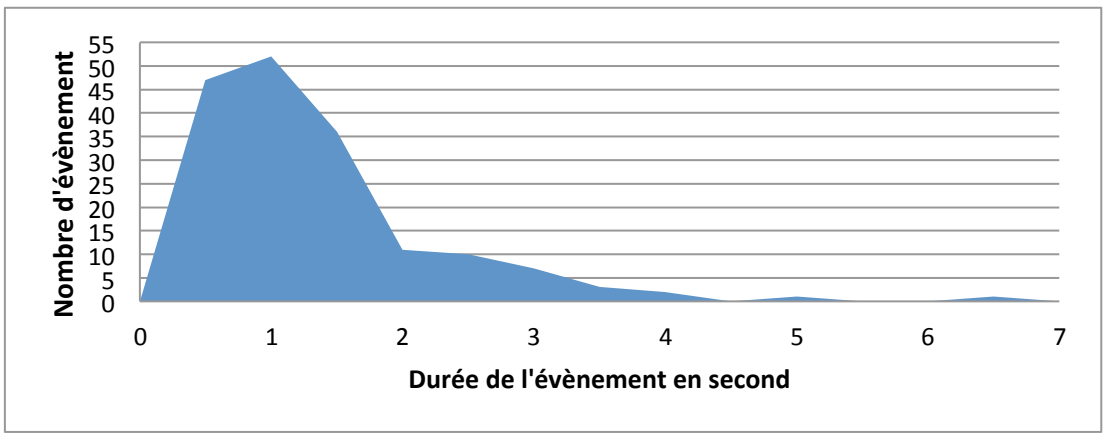

Figure 20 : Répartition de nombre d'évènement de rire dans l'annotation audio de la sous-partie du corpus ROMEO2 des 15 sujets utilisés pour l'expérimentation du système automatique. Le nombre d'évènement est calculé par un intervalle de durée de 0,5 second.

Alors que l'annotation vidéo se focalise sur le signe visuel de rire et sourire, en revanche, l'annotation audio se concentre sur le signe auditif, la corrélation entre les

annotations audio et vidéo pour ces 15 sujets est donc intéressante à étudier afin de chercher un lien entre les indices. Parmi les 170 annotations audio, il existe 100 annotations audio dont la période est incluse dans une annotation vidéo de rire ou sourire avec bouche ouverte, 3 annotations audio dont la période couvre une annotation vidéo de rire ou sourire avec bouche ouverte, 49 annotations audio qui ont une intersection avec une annotation vidéo de rire ou sourire avec bouche ouverte. Le ratio moyen de superposition au niveau de la durée de ces 49 annotations audio est de 71,83%. De plus, il existe également 18 annotations audio qui n'ont pas d'intersection avec l'annotation vidéo. J'ai vérifié ces annotations et mis la cause dans le Figure 22.

Tableau 28 : la cause de laquelle les annotations audio n'ont pas d'intersection avec une annotation vidéo.

Cause	Nb d'évènement
Signe visuel trop court	8
Signe visuel pas évident	4
Visage caché	3
Audio annoté hors la période de l'expérimentation	2
Rire de l'expérimentateur	1
Total	18

7.4 Système de la détection visuelle

7.4.1 Protocoles de l'évaluation sur le corpus ROMEO2

La séparation de donnée d'apprentissage et de donnée à tester suit trois protocoles différents :

- **Mono-sujet**: pour chaque sujet, la moitié des données est utilisée pour l'entrainement du modèle statistique, et l'autre moitié des données du sujet est pour le test.
- **Multi-sujet**: pour chaque sujet, la moitié des données ainsi que les données de tous les autres sujets sont utilisées pour l'entrainement du modèle statistique, et l'autre moitié de données restantes du sujet est pour le test.
- «**Leave-one-out**»: pour chaque sujet, l'entrainement du modèle statistique est effectué en utilisant les données de tous les autres sujets, et le test se fait sur toutes les données du sujet.

Afin de réduire la charge de calcul et d'équilibrer la quantité d'image entre les frames de sourire/rire et les frames non-sourire/non-rire, un protocole de sélection a été mis en place en utilisant les données d'annotation. Pour les frames non-sourire et non-rire, nous avons pris la première image d'un visage détecté chaque 1/6ème de seconde. Le

nombre de frames de sourire et rire a été également réduit en utilisant le même protocole avec 1/15ème de seconde d'intervalle. Les images sélectionnées sont ensuite transmises au système d'entrainement statistique pour la classification.

L'évaluation du système vidéo est effectuée au niveau du frame, du segment et en continu « In the wild ». Notre évaluation « in the wild » est différente de celle du workshop EmotiW 2014 [43]. Dans notre expérimentation, nous considérons une évaluation « in the wild » comme l'évaluation en flux continu dans un corpus réaliste non segmenté en prenant compte les variabilités de l'environnement, de la condition de la luminosité, de la pose de la tête, etc. Pour l'évaluation au niveau du segment, nous considérons qu'un segment contient un sourire ou un rire si le ratio de frames classées comme sourire ou rire qu'il contient dépasse un certain seuil. L'évaluation en continu prend une décision tous les 100ms en considérant les résultats individuels de détection des trois frames incluses dans ces 100ms.

7.4.2 Résultat au niveau de frame et segment

Un total de 389 évènements de sourire et rire annotés est utilisé dans nos expériences. Le système visuel envisage principalement deux types de problèmes : l'échec à trouver le visage en raison d'un problème de détection visuelle (souvent causé par la rotation de la tête des sujets) et la ressemblance apparente entre une bouche parlante et une bouche ouverte souriante causée par les rides du visage. Afin d'évaluer précisément la performance du système de détection visuelle, nous supposons l'utilisation d'un système de détection d'activité vocale avec des sorties parfaites. Par conséquent, tous les segments qui contiennent la parole du sujet sont considérés comme des segments de non rires et non sourire. Le sous-ensemble du corpus ROMEO2 utilisé dans nos expériences contient 13835 secondes, dont 4984 secondes sont le discours des sujets. La proportion de frames que le système visuel peut détecter est montrée dans la Figure 21.

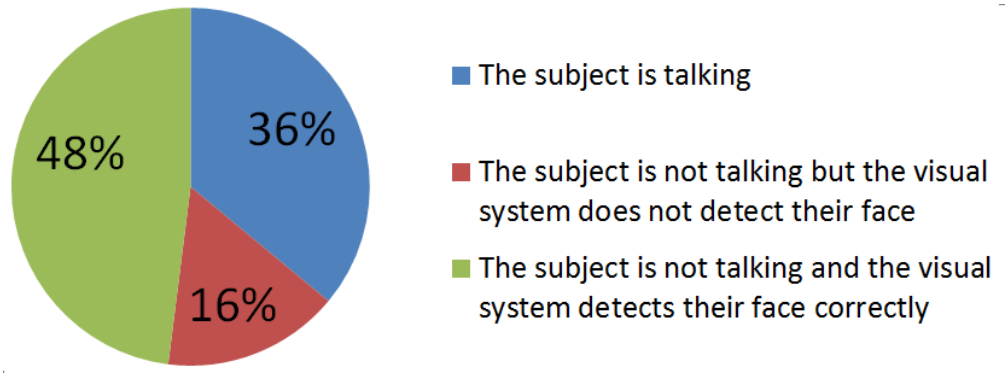

Figure 21 : proportion de frames que le système visuel peut détecter

Sur les 389 évènements de rire et sourire annotés, le visage des sujets est correctement suivi pour 292 événements. Ce problème limite le meilleur rappel de

reconnaissance de segment de rire et sourire du système visuel à 292 sur 389 au lieu de 100 pourcents à l'origine, donc le rappel maximal que le système visuel peut attendre est d'environ 75%. Nous avons également extrait 378 segments aléatoires en dehors des segments de rire et sourire annotés pour calculer le taux de bonne reconnaissance et la précision du système visuel au niveau de l'évaluation segmentale. L'évaluation segmentale du système visuel a été donc calculée à la base de l'ensemble de ces 767 segments.

Nous considérons qu'un segment contient un sourire ou un rire si le ratio de frames classés comme sourire ou rire qu'il contient dépasse le seuil. Le résultat de l'évaluation segmentale du système visuel sous les trois protocoles est illustré dans la Figure 22, le ratio du seuil est dans un intervalle de 0 à 70%. Le système sous le protocole mono-sujet atteint sa F-mesure optimale avec le seuil 15%, la performance correspondante est composée par un rappel de 65,8%, une précision de 81,9% et une F-mesure de 73,0%. Le système sous le protocole multi-sujet atteint sa F-mesure optimale avec le seuil 20%, la performance correspondante est composée par un rappel de 62,5%, une précision de 83,9% et une F-mesure de 71,6%. Par contre, la F-mesure du système sous le protocole de « leave-one-out » diminue quand le seuil augmente en raison de la forte décroissance du taux de rappel, sa F-mesure optimale est avec le seuil 0.

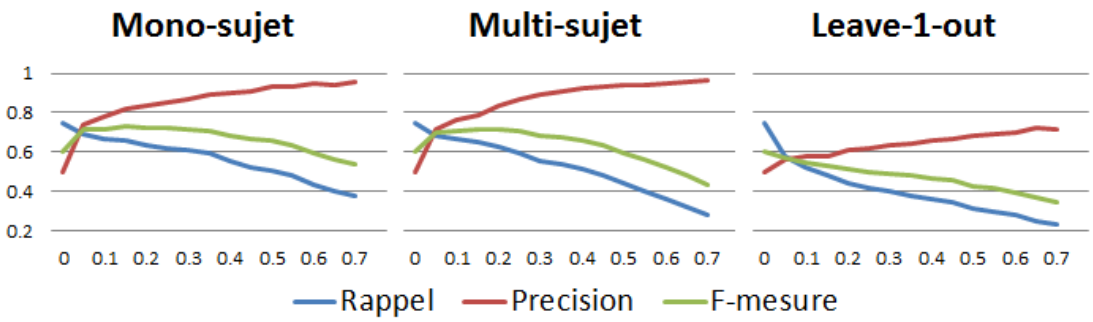

Figure 22 : Evaluation segmentale du système visuel sous trois protocoles de séparation de données d'apprentissage et de teste avec un intervalle de ratio de seuil de validation segmentale au niveau de nombre de frame entre 0 et 70%.

Tableau 29 : Matrice de confusion pour l'évaluation du système visuel au niveau de frame sous le protocole mono-sujet. Les valeurs dans le tableau correspondent au pourcentage du nombre de frame par rapport au nombre total de frame.

		Prédiction	
		Rire/sourire	**Non-rire/non-sourire**
Annotation	**Rire/sourire**	14,5%	3,5%
	Non-rire/non-sourire	10,8%	71,2%

Le résultat de l'évaluation segmentale des systèmes sous trois différents protocoles avec un seuil commun de 20% est montré dans la colonne de droite du Tableau 30. Le résultat de l'évaluation au niveau du frame tout au long de l'expérimentation avec les

15 sujets est montré dans la colonne du milieu du Tableau 30. Comme le nombre de frame de sourire et rire n'est pas équilibré avec le nombre de frame de non-rire/non-sourire (voir la matrice de confusion dans le Tableau 29), l'évaluation au niveau de frame utilise le ratio de bonne reconnaissance et le ratio d'erreur balancé (« Balanced Error Rate » en anglais) pour présenter les résultats.

A partir du Tableau 30, nous pouvons observer que la meilleure performance est atteinte en utilisant le protocole dépendant du sujet (soit mono-sujet) donc sans données supplémentaires provenant d'autres sujets pour l'entrainement du modèle statistique. Lorsque nous ajoutons des données des autres sujets pour l'entrainement, nous constatons une baisse de performance du système. Ceci peut être expliqué par le fait que, pour les personnes âgées, le visage ainsi que les expressions de sourire et de rire sont sensiblement différents d'une personne à une autre. Enfin, le protocole « leave-one-out » (aucune donnée du sujet de test utilisée pour entrainer le modèle) a donné la plus faible performance parmi les trois protocoles.

Tableau 30 : Evaluation du système visuel sous les trois protocoles. La colonne du milieu correspond à l'évaluation au niveau de la frame. La colonne de droite correspond à l'évaluation segmentale avec un seuil commun de 20%. A. signifie le taux de bonne reconnaissance (« accurancy » en anglais), BER. signifie le ratio d'erreur balancé, R. signifie le rappel, P. signifie la précision et F signifie la F-mesure.

Protocole (%)	Evaluation au niveau du frame	Evaluation segmentale (seuil 20%)
Mono-sujet	A. 85,7 R. 80,6 BER. 23,7	A. 84,0 R. 63,8 P. 83,4 F. 72,3
Multi-sujet	A. 73,9 R. 62,4 BER. 31,4	A. 83,6 R. 62,5 P. 83,9 F. 71,6
Leave-one-out	A. 72,8 R. 40,8 BER. 46,6	A. 60,5 R. 44,2 P. 60,9 F. 51,2

7.4.3 Résultat du système visuel « in the wild »

La séparation des données d'apprentissage et de test pour le système « in the wild » utilise le protocole « leave-one-out » pour évaluer la performance du système indépendamment des sujets. Le système visuel « in the wild » prend la décision de détection de rire et sourire toutes les 100 ms en considérant les résultats de détection des 3 frames correspondantes (le signal vidéo d'entrée du système est enregistré en 30 frames par second, d'où 3 frames pour 100 ms). Si au moins la moitié de frames où le visage du sujet peut être reconnu sont classifiés en rire ou sourire, la fenêtre de période de 100 ms sera classifiée comme rire ou sourire.

Un filtre d'analyse est également appliqué sur les résultats des fenêtres de 100 ms. Le filtre analyse les 8 résultats de 100 ms autour de l'instant T (donc entre T-400 ms et T+400 ms), et fournit une décision positive si le nombre de fenêtres de 100 ms classifiées comme rire dans ces 800 ms dépasse le seuil. L'évaluation du système visuel par le ratio d'erreur balancé (BER) en fonction du seuil de filtre d'analyse testé en prenant l'annotation vidéo comme l'annotation de référence est illustrée dans le schéma de gauche de la Figure 23. Pour comparer avec la performance du système de fusion audio-visuelle, l'évaluation du système visuel en prenant l'ensemble des

annotations audio et vidéo comme annotation de référence est également testée et montrée dans le schéma de droite de la Figure 23. Nous pouvons voir que, les ratios d'erreur balancés dans les deux tests arrivent à leur valeur minimale en prenant le seuil 3.

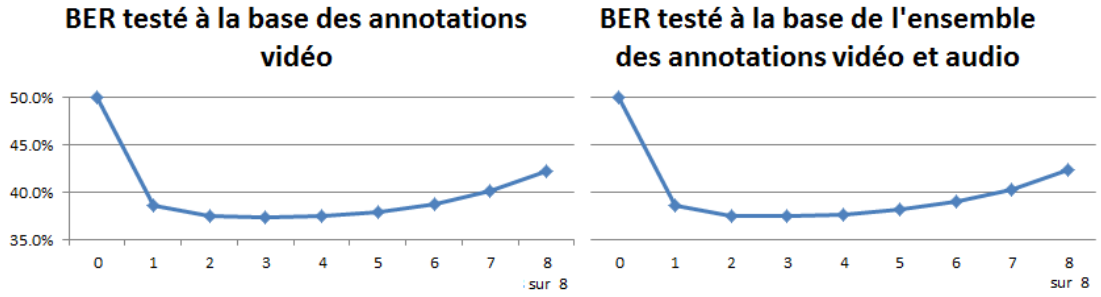

Figure 23 : Ratio d'erreur balancé (BER) en fonction du seuil de filtre d'analysé

Le Tableau 31 résume l'évaluation de la performance du système visuel « in the wild » avec ou sans le filtre d'analyse en prenant l'annotation vidéo ou l'ensemble des annotations vidéo et audio comme annotation de référence. Nous pouvons voir que, le filtre d'analyse améliore dans tous les cas le rappel et le ratio d'erreur balancé. De plus, comme la plupart des annotations audio de rire sont partiellement ou totalement incluses dans les annotations vidéo, les résultats de l'évaluation utilisant les deux annotations comme référence se ressemblent beaucoup.

Tableau 31 : Evaluation du système visuel « in the wild » avec ou sans le filtre d'analyse en prenant l'annotation vidéo ou l'ensemble des annotations vidéo + audio comme annotation de référence. A. signifie le taux de bonne reconnaissance, R. signifie le rappel. BER. signifie le ratio d'erreur balancé.

Unité : %		Choix de l'annotation de référence	
		Annotation vidéo	Annotation vidéo + audio
Filtre d'analyse	Seuil 3 sur 8	A. 69,1 R. 54,6 BER. 37,3	A. 69,0 R. 54,3 BER. 37,4
	Sans	A. 69,5 R. 50,2 BER. 45,1	A. 69,5 R. 50,6 BER. 44,9

7.5 Système de la détection auditive

7.5.1 Résultat à l'évaluation segmentale

La méthode de la classification auditive utilisée pour l'évaluation segmentale est faite au niveau de la trame à l'aide des 13 coefficients de MFCC et de 4 modèles GMM afin de représenter ces 4 classes (parole du sujet, rire, parole du robot, silence). Chaque trame a une longueur de 20 ms avec un décalage de 10 ms de la trame précédente.

Comme ce qui est montré dans la sous-section « 7.3.2 Sous-partie du Corpus ROMEO2 », la plupart des annotations audio sont partiellement ou totalement incluses dans les annotations vidéo, l'évaluation segmentale du système auditif prend

l'annotation vidéo comme annotation de référence vidéo. Cela facile la comparaison avec l'évaluation segmentale du système visuel ainsi que la fusion des systèmes des deux modalités. Par l'étude des annotations du corpus ROMEO2, nous avons remarqué que, visuellement, le signe visuel du rire dure plus longtemps que son signal audio respectif. Le signe auditif du rire perçu peut effectivement être très bref en comparaison des signes du visage ou du corps. Il peut être partiellement ou totalement masqué par la parole du robot ou fusionné avec la parole du sujet et/ou le silence. Par conséquent, pour l'évaluation segmentale des annotations vidéo par la détection auditive illustrée dans la Figure 24, nous avons utilisé une stratégie d'agrégation qui s'applique sur de courtes séquences de trames consécutives (une fenêtre glissante de 800 ms par exemple) à la place de l'ensemble de la tranche audio alignée avec une annotation visuelle. L'objectif est de détecter la présence d'événements brefs de rires en audio dans une longue annotation de rire perçue visuellement par l'humain.

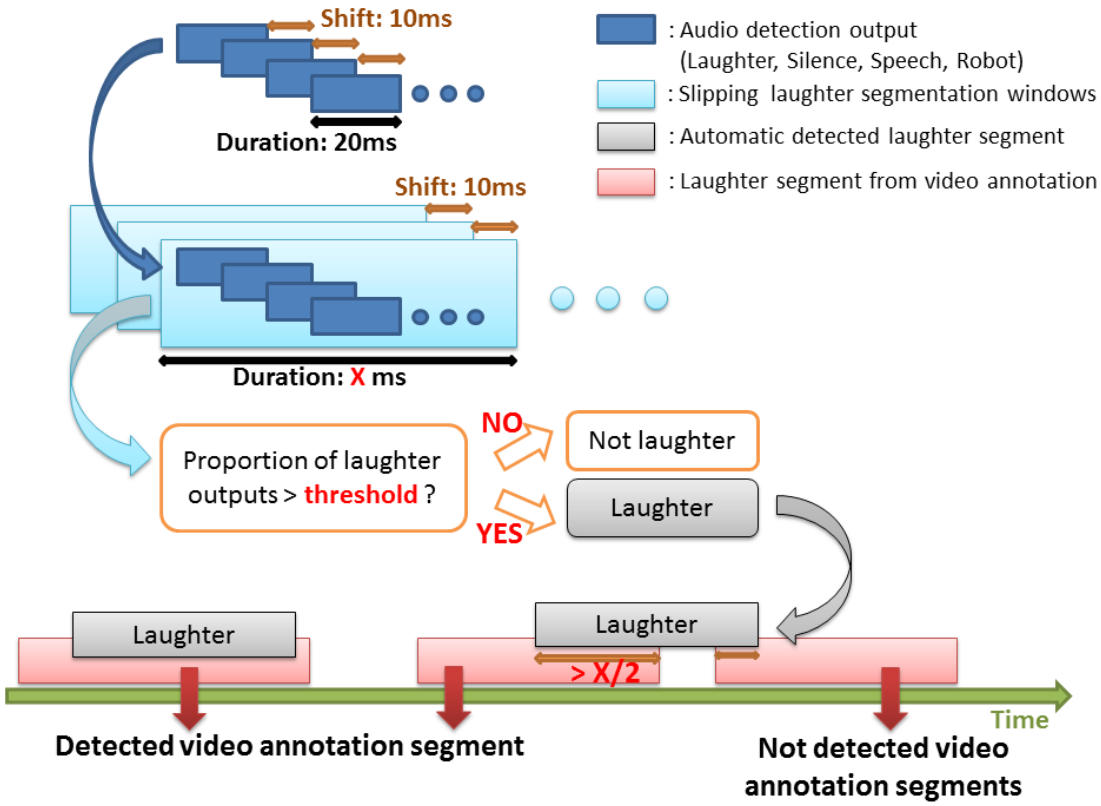

Figure 24 : schéma de l'évaluation segmentale du système de la détection auditive de rire

Pour la détection de rires, nous avons également 389 événements de sourire et de rire annotés ainsi que 378 segments extraits au hasard dans les périodes qui ne contiennent ni sourire ni rire. Un événement de rire est considéré comme classé correctement par le système si au moins une fenêtre d'analyse glissante est reconnue comme un rire par le système. Un segment non rire est inversement considéré comme correctement classés si aucune fenêtre d'analyse glissante dedans n'est reconnue comme le rire. En outre, comme chaque fenêtre d'analyse représente une séquence de trames, deux

stratégies d'agrégation sont utilisées : par le vote de majorité et par le seuil. Pour la première stratégie, la fenêtre reçoit l'étiquette de la classe la plus représentée (parmi les 4 classes : parole du sujet, rire, parole du robot, silence). Quant à la seconde, le ratio d'occurrences d'une classe dans la fenêtre doit être supérieur au seuil pour que la fenêtre soit marquée avec l'étiquette de cette classe.

Tableau 32 : Evaluation du système auditif de la détection de rire en utilisant des fenêtres d'analyse de différentes durées sur 389 événements de rire et sourire et 378 segments non-rire et non-sourire. Vote Maj.: une fenêtre est considérée comme le rire si la classe la plus représentée des trames dans la fenêtre est le rire. Seuil 50%: une fenêtre doit contenir au moins 50% des trames de rire pour être considéré comme un rire. A. signifie le taux de bonne reconnaissance, R. signifie le rappel, P. signifie la précision et F signifie la F-mesure.

Duré de fenêtre (%)	Vote Maj.	Seuil 50%
600 ms	A. 55,9 R. 80,7 P. 54,4 F. 65,0	A. 57,8 R. 65,6 P. 57,3 F. 61,2
800 ms	A. 56,1 R. 75,8 P. 54,8 F. 63,6	A. 58,4 R. 57,1 P. 59,4 F. 58,2
1 s	A. 57,2 R. 71,7 P. 56,1 F. 63,0	A. 57,9 R. 50,4 P. 60,1 F. 54,8

Le Tableau 32 montre les résultats obtenus par le système. Nous pouvons voir, d'une part, qu'une courte fenêtre ramène un meilleur rappel mais une précision relativement faible alors qu'une longue fenêtre conduit au résultat inverse. D'autre part, une stratégie d'agrégation plus rigoureuse (avec un seuil au moins de 50%) améliore la précision au détriment du rappel.

Notez que sur les 389 évènements de rire et sourire dans l'annotation vidéo, seulement 168 sont des rires. Par conséquent, lorsque nous lançons le système auditif sur ces 168 événements de rire, en utilisant une fenêtre d'analyse de 800 ms, on obtient respectivement un rappel de 85,7% et 69,6% pour la stratégie du vote à la majorité et la stratégie du seuil de 50%, qui surpasse significativement la performance de la détection de l'ensemble de rire et sourire.

7.5.2 Résultat du système auditif « in the wild »

La méthode de la classification auditive utilisée dans le système « in the wild » est différente de celle présentée dans la partie de l'évaluation segmentale. Comme la durée d'un rire est généralement supérieure à 800ms, une fenêtre glissante de 800ms qui calcule la classe dominante qui possède la plus grande quantité de séquence de 20ms est utilisée comme la sortie initiale du système audio. La sortie du système est donc composée par les résultats de classification d'une fenêtre glissante de 800 ms avec un décalage de 100 ms. Afin de corréler avec la sortie du système visuel « in the wild », la sortie du système est convertie afin de prendre la décision de détection auditive de rire pour une suite de fenêtres de 100 ms décalées (durée 100 ms, décalage 100 ms). Comme chaque fenêtre de 100 ms est couverte par 8 fenêtres glissantes de 800 ms de la sortie du système à l'origine, nous utilisons également les deux

stratégies d'agrégation mentionnées dans l'évaluation segmentale : par le vote majoritaire et par le seuil.

Tableau 33 : Evaluation du système auditif « in the wild » en prenant l'annotation audio ou l'ensemble des annotations vidéo + audio comme annotation de référence. Vote Maj.: une fenêtre de 100 ms est considérée comme le rire si la classe la plus représentée des 8 fenêtres de 800 ms couvrant la fenêtre de 100 ms est le rire. Seuil 50%: une fenêtre de 100 ms est considérée comme le rire si la majorité des 8 fenêtres de 800 ms couvrant la fenêtre de 100 ms est le rire. A. signifie le taux de bonne reconnaissance, R. signifie le rappel. BER. signifie le ratio d'erreur balancé.

(Unité : %)	Choix de référence d'annotation	
Stratégies d'agrégation	Annotation audio	Annotation audio + vidéo
Seuil 50%	A. 90,8 R. 58,7 BER. 45,5	A. 83,8 R. 19,2 BER. 43,8
Vote Maj.	A. 89,3 R. 60,4 BER. 46,0	A. 82,7 R. 21,2 BER. 44,3

Le résultat de l'évaluation du système auditif en utilisant l'annotation audio comme annotation de référence est illustré dans la colonne du milieu du Tableau 33. Pour comparer avec la performance du système de fusion audio-visuelle, l'évaluation du système visuel en prenant l'ensemble des annotations audio et vidéo comme annotation de référence est montrée dans la colonne de droite du Tableau 33. Nous pouvons voir que, la stratégie d'agrégation plus rigoureuse (avec un seuil d'au moins 50%) améliore le ratio d'erreur balancé au détriment du rappel. Le rappel sur l'annotation audio est beaucoup élevé que celui sur l'ensemble des annotations audio et vidéo. L'utilisation de l'annotation vidéo diminue le rappel parce que le système auditif n'est pas capable de détecter les sourires dans l'annotation vidéo, en revanche, le ratio d'erreur balancé est amélioré en raison de la baisse du nombre de fausse alarme.

7.6 Système audio-visuel

7.6.1 Résultat de l'évaluation segmentale

Dans notre système de détection à la base de la fusion de décision entre les systèmes audio et vidéo, nous considérons une bonne détection de rire et sourire si l'une des deux modalités donne une détection positive. Pour faire un compromis entre la précision et le rappel dans le système audio, nous utilisons une fenêtre d'analyse de 800 ms. Nous testons également les deux stratégies d'agrégation de trames audio mentionnées dans la section de l'évaluation du système auditif. Les trois protocoles de séparation de données d'apprentissage et de test au niveau du système de détection visuelle sont également testés en prenant un intervalle de ratio de seuil de validation segmentale au niveau du nombre de frame entre 0 et 70%. La performance des systèmes de fusion est illustrée dans la Figure 25. Nous pouvons voir que, la performance des systèmes de fusion varie doucement en fonction du seuil de validation segmentale du système visuel. Les systèmes sous le protocole mono-sujet et multi-sujet atteignent leur F-mesure optimale autour d'une valeur de seuil 20%.

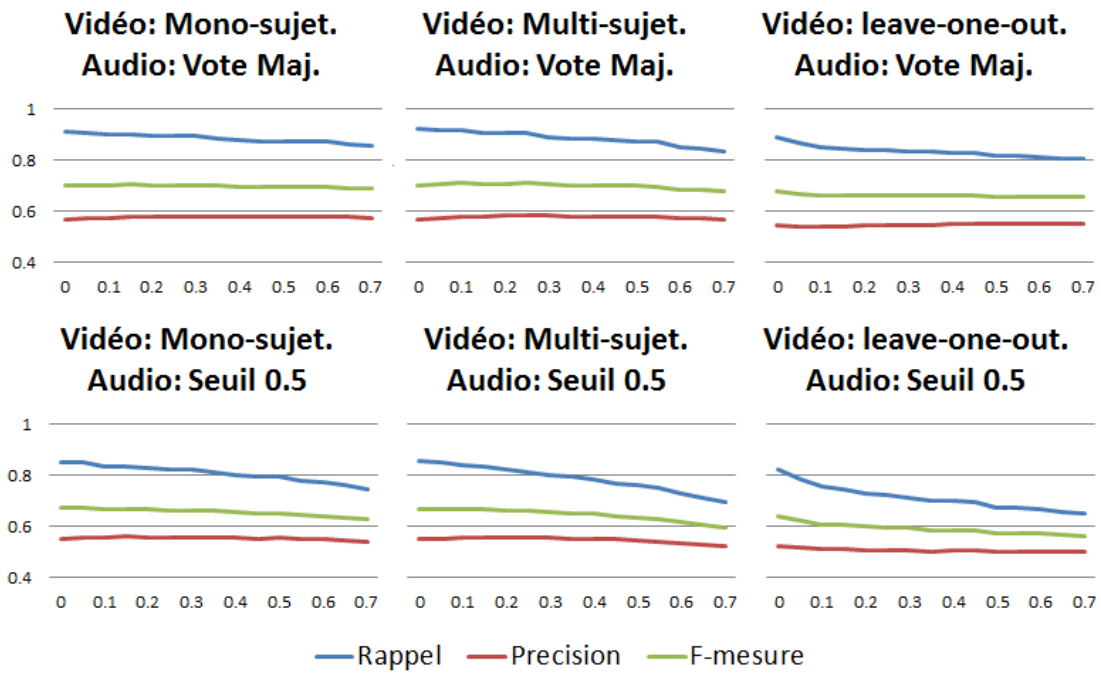

Figure 25 : Performance à l'évaluation segmentale des systèmes de fusion audio-visuelle en combinant les deux stratégies d'agrégation de trames pour le système audio et les trois protocoles de séparation de données d'apprentissage et de test pour le système visuel avec un intervalle de ratio de seuil de validation segmentale au niveau de nombre de frame entre 0 et 70%..

Tableau 34 : Evaluation du système de détection audio-visuelle de rire et de sourire sur 389 évènements de rire et sourire et 378 segments non-rire et non-sourire. Pour la détection audio, une fenêtre d'analyse de 800 ms est utilisée. Pour la détection vidéo, le seuil prend la valeur 20%. A. signifie le taux de bonne reconnaissance, R. signifie le rappel, P. signifie la précision et F signifie la F-mesure.

		Fusion audio-visuelle	
Protocole (Unité : %)	**Vidéo seul**	**Avec Audio Vote Maj.**	**Avec Audio Seuil 50%**
Mono-sujet	A. 84,0 R. 63,8 P. 83,4 F. 72,3	A. 61,4 R. 89,5 P. 57,7 F. 70,2	A. 57,9 R. 82,5 P. 55,7 F. 66,5
Multi-sujet	A. 83,6 R. 62,5 P. 83,9 F. 71,6	A .62,1 R. 90,5 P. 58,1 F. 70,8	A. 57,8 R. 82,0 P. 55,7 F. 66,3
« Leave-one-out »	A. 60,5 R. 44,2 P. 60,9 F. 51,2	A. 55,9 R. 83,8 P. 54,2 F. 65,9	A. 50,3 R. 72,8 P. 50,7 F. 59,8

Le Tableau 34 montre les résultats obtenus des systèmes de fusion avec un seuil de 20%. Nous pouvons voir que, la fusion ramène une forte amélioration du rappel mais une diminution de la précision et du ratio de bonne connaissance par rapport au système visuel tout seul. Au niveau de la F-mesure, les systèmes de fusion sous les protocoles dépendant du sujet (mono-sujet et multi-sujet) dans la partie de détection visuelle n'ont pas réussi d'améliorer la F-mesure du système visuel tout seul. La F-mesure du système de fusion reste cependant entre celle des deux systèmes de mono-modalité et meilleure que celle du système auditif. Cela peut être dû à l'utilisation unifiée du protocole indépendant du sujet (« leave-one-out ») dans la partie de détection auditive. La faible précision en utilisant le protocole indépendant du sujet

dans le système audio augmente beaucoup le nombre de fausses alarmes pour le système de fusion. Dans le cas où les systèmes des deux modalités utilisent tous les deux le protocole « leave-one-out », la F-mesure du système de fusion est bien meilleure que celle des deux systèmes de mono-modalité.

7.6.2 Résultat à l'évaluation « in the wild »

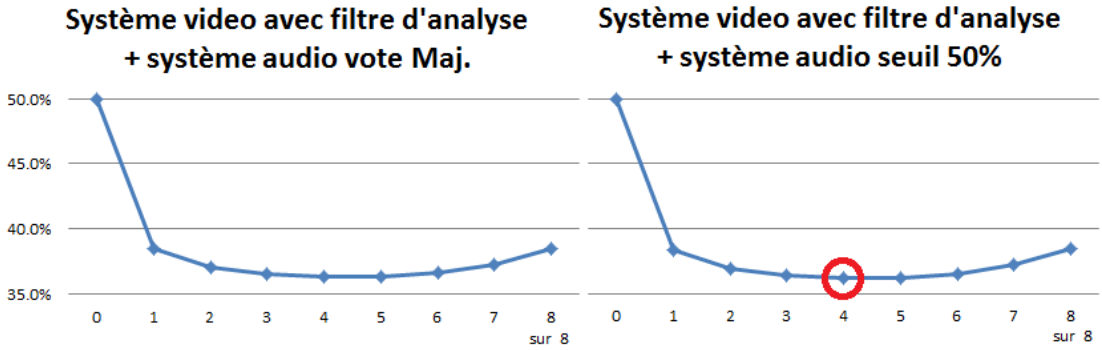

Figure 26 : Ratio d'erreur balancé du système de fusion directe en fonction du seuil du filtre d'analyse dans le partie vidéo et suivant les deux stratégies d'agrégation pour la partie audio.

Dans le système audio-visuel « in the wild », deux types de fusion sont testés. Premièrement, nous fusionnons simplement les sorties des 2 systèmes et les comparons avec l'ensemble de l'annotation audio et vidéo. Nous considérons une bonne détection si au moins l'une des modalités détecte un évènement de rire ou sourire dans la fenêtre de 100 ms, et cette période de 100 ms est incluse dans au moins l'une des annotations audio et vidéo. Le filtre d'analyse de 800 ms pour le système visuel est appliqué. Le ratio d'erreur balancé du système de fusion directe en fonction du seuil du filtre d'analyse dans la partie vidéo et suivant les deux stratégies d'agrégation pour la partie audio est illustré dans Figure 26. Le meilleur ratio d'erreur balancé est marqué par le rond rouge.

Le Tableau 35 montre la performance des systèmes de fusion directe audio-visuelle. Le résultat montre que la fusion directe du système auditif et visuel amène une amélioration du rappel et du ratio d'erreur balancé par rapport au système monomodal.

Tableau 35 : Evaluation du système de fusion directe audio-visuelle « in the wild » en simplement fusionnant la sortie de système auditif et visuel. A. signifie le taux de bonne reconnaissance, R. signifie le rappel. BER. signifie le ratio d'erreur balancé.

(Unité : %)	Fusion audio-visuelle	
Vidéo seul (seuil 4 sur 8)	Avec Audio Vote Maj.	Avec Audio Seuil 50%
A. 73,5 R. 48,1 BER. 37,6	A. 67,6 R. 58,7 BER. 36,3	A. 68,5 R. 57,8 BER. 36,2

Le deuxième type de fusion ajoute une couche cascade en utilisant la sortie de la classe « parole du sujet » du système auditif avant l'application du système visuel, c'est-à-dire que le système visuel est désactivé si le système auditif classifie la fenêtre

de 100 ms comme de la parole du sujet. Nous espérons que la détection de parole permet d'éviter l'influence du mouvement de la bouche parlante et de réduire les faux positifs dans la détection visuelle de rire et sourire. Par contre, le ratio d'erreur balancé du système de fusion en cascade est de 38,2% pour l'audio en vote majoritaire et de 37,9% pour l'audio avec un seuil 50%. Le système de fusion en cascade ne fournit pas donc d'amélioration pour le système de fusion audio-visuelle pour la détection de rire ou sourire « in the wild ».

7.7 Corrélations statistiques entre la performance des systèmes et les questionnaires et les annotations

7.7.1 Corrélations liées à la performance de l'évaluation segmentale

Tableau 36 : Corrélation de la performance de l'évaluation segmentale du système visuel sous le protocole mono-sujet avec un seuil de validation segmentale de 20%. P.signifie la valeur P. Cor. signifie la valeur de corrélation.

Classe 1	Classe 2	P.	Cor.
Ratio de bonne reconnaissance	Conscienciosité	0,015	-0,612
	Ouverture	0,022	-0,586
	(Q9) Posséder un robot?	0,013	-0,624
Rappel	(Q6) Nom humain (1) ou non-humain (0)	0,050	-0,515
	(Q9) Posséder un robot?	0,017	-0,606
Précision	Agréabilité	0,020	-0,593
	Extraversion	0,025	0,575
	(Q3) Nao est agréable de vous?	0,053	-0,508
F-mesure	(Q9) Posséder un robot?	0,024	-0,579

Les corrélations de la performance de l'évaluation segmentale du système visuel pour les 15 personnes âgées testées sous le protocole mono-sujet, multi-sujet et « leave-one-out » sont montrées dans le Tableau 36, le 错误！书签自引用无效。 et le Tableau 38 respectivement. Celle du système audio-visuel est montrée dans le Tableau 39. Nous pouvons conclure que, parmi les cinq dimensions de personnalité, la conscienciosité, l'ouverture, l'agréabilité ont un effet négatif sur la performance de la détection du système visuel pour les personnes âgées; en revanche, l'extraversion et la stabilité émotionnelle ont un effet positif sur la performance du système. Nous pouvons également noter que, plus le robot agit comme un humain agréable et plus les personnes âgées sont intéressées par l'expérimentation, moins la performance obtenue est fiable. Ce phénomène peut être expliqué par le fait que les personnes âgées qui sont plus intéressées par l'expérimentation, rient plus souvent (ce qui est démontré dans le Tableau 15) et sont plus à l'aise donc avec des poses variées. La variété des poses et le manque de données correspondantes amènent des difficultés pour la

reconnaissance de rire et de sourire avec le système statistique visuel. De plus, nous observons que, la performance du système visuel est également influencée par le nombre de rotation de la tête du sujet.

Tableau 37 : Corrélation de la performance de l'évaluation segmentale du système visuel sous le protocole multi-sujet avec un seuil de validation segmentale de 20%.

Classe 1	Classe 2	P.	Cor.
Ratio de bonne reconnaissance	Ouverture	0,027	-0,569
	(Q11) Considérer le robot comme machine (1) ou humain (0)?	0,015	0,613
	(Q9) Posséder un robot?	0,035	-0,546
Rappel	Ouverture	0,048	-0,517
Précision	Stabilité émotionnelle	0,035	0,546
	(Q4) Nao est poli?	0,045	0,525
	(Q11) Considérer le robot comme machine (1) ou humain (0)?	0,048	0,518
F-mesure	Nombre de rotation de la tête	0,049	-0,516
	Ouverture	0,030	-0,560
	(Q11) Considérer le robot comme machine (1) ou humain (0)?	0,022	0,585
	(Q9) Posséder un robot?	0,041	-0,533

Tableau 38 : Corrélation de la performance de l'évaluation segmentale du système visuel sous le protocole « leave-one-out » avec un seuil de validation segmentale de 20%.

Classe 1	Classe 2	P.	Cor.
Ratio de bonne reconnaissance	(Q9) Posséder un robot?	0,045	0,525
Précision	Nombre de rotation de la tête	0,016	-0,607

Tableau 39 : Corrélation de la performance de l'évaluation segmentale du système audio-visuel sous le protocole « leave-one-out ».

Classe 1	Classe 2	P.	Cor.
Ratio de bonne reconnaissance	(Q4) Nao est poli?	0,032	-0,554

7.7.2 Corrélations liées à la performance de l'évaluation « in the wild »

Les corrélations de la performance de l'évaluation « in the wild » du système audio, vidéo et audio-visuel pour les 15 personnes âgées testées sont montrées dans le Tableau 40, le Tableau 41 et le Tableau 42 respectivement. Le Tableau 40 et le Tableau 42 montrent une corrélation négative entre la performance et le nombre d'évènement de rire et sourire pour le système visuel et le système audio-visuel, cette corrélation correspond au phénomène présenté dans la sous-section précédente sur l'évaluation segmentale du système visuel. Le Tableau 41 montre que le système

auditif subit également une baisse de performance quand les personnes âgées sont plus intéressées par l'expérimentation.

Tableau 40 : Corrélation de la performance du système visuel « in the wild » avec le nombre d'évènement du rire et sourire dans l'annotation vidéo.

Classe 1	Classe 2	P.	Cor.
Ratio de bonne reconnaissance	Nb d'évènement rire et sourire	0,014	-0,618
Ratio d'erreur balancé		0,024	0,578

Tableau 41 : Corrélation de la performance du système auditif « in the wild » avec le questionnaire de satisfaction. L'évaluation utilise l'annotation audio comme annotation de référence.

Classe 1	Classe 2	P.	Cor.
Ratio de bonne reconnaissance	(Q11) Considérer le robot comme machine (1) ou humain (0)?	0,0513	0,511
	(Q10) Préférer une apparence robot (0) ou un humain (1)?	0,0027	-0,716
Ratio d'erreur balancé	(Q3) Nao est agréable de vous?	0,0368	0,542
	(Q11) Considérer le robot comme machine (1) ou humain (0)?	0,0149	-0,614

Tableau 42 : Corrélation de la performance du système audio-visuel « in the wild » avec le nombre d'évènement de rire et sourire dans l'ensemble des annotations audio et vidéo et le questionnaire de satisfaction.

Classe 1	Classe 2	P.	Cor.
Ratio de bonne reconnaissance	Nb d'évènement rire et sourire	0,0076	-0,659

7.8 Conclusion partielle

Ce chapitre de ma thèse présente un système de détection audio-visuelle de sourire et rire dans le contexte de l'interaction sociale entre les personnes âgées et un robot. La fusion directe des systèmes audio et vidéo se produit en deux étapes: d'abord, chaque système de modalité est exécuté séparément pour produire sa décision individuelle sur un segment aligné des signaux audio et vidéo. La décision finale est positive (il y a un rire ou un sourire) si au moins l'un des deux systèmes délivre une décision positive.

Le système visuel donne une bonne précision au-dessus de 80% avec un rappel au-dessus de 60% au niveau de l'évaluation segmentale dépendante du sujet et une précision autour de 60% dans les évaluations segmentale indépendantes du sujet (« leave-one-out »). Il a cependant un rappel relativement faible en comparaison avec le système audio sous le protocole de vote majoritaire au niveau de l'évaluation segmentale. Cela peut être dû au fait que de nombreux sujets rient en tournant la tête ou à l'instant où la détection de visage ne pouvait pas détecter le visage. Pendant ce temps, le système audio fonctionne en continu ce qui conduit à un taux de rappel relativement élevé. Nous croyons que le rappel du système visuel peut être amélioré en traitant la question de la rotation de la tête alors que le système audio peut

entraîner une meilleure précision en utilisant plus de données pour l'entrainement du modèle statistique. Quand les deux systèmes monomodaux utilisent tous le protocole de séparation de données d'apprentissage et de test « leave-one-out », la fusion des deux systèmes améliore significativement la performance d'un système monomodal au niveau de l'évaluation segmentale. L'amélioration du système de chaque modalité ramènera en effet une amélioration globale de la performance du système de fusion.

Le système de détection audio-visuelle « in the wild » est également étudié dans ce chapitre en appliquant le protocole « leave-one-out » dans la partie vidéo ainsi que la partie audio, toutes les difficultés dues à l'âge des sujets et au corpus d'interaction sociale réaliste augmentent le défi de la détection. La fusion directe des sorties des deux systèmes amènent une forte amélioration du rappel et une petite amélioration du ratio d'erreur balancée. L'utilisation de la détection auditive de parole est également testée dans le système de fusion pour éviter l'influence du mouvement de la bouche parlante, mais le résultat ne montre pas d'avantage par rapport à la fusion directe.

Les corrélations entre la performance des systèmes et les questionnaires et les annotations pour les 15 personnes âgées testées sont également analysées. Nous avons trouvé une corrélation négative entre la performance et le nombre d'évènements de rire et sourire pour le système visuel et le système audio-visuel. De plus, les systèmes audio, vidéo et audio-visuel subissent également une baisse de performance quand les personnes âgées sont plus intéressées par l'expérimentation. Ce phénomène peut être expliqué par l'effet que les personnes âgées qui sont plus intéressées par l'expérimentation rient plus souvent et sont plus à l'aise donc avec des poses variées. La variété des poses et le manque de données correspondantes amènent des difficultés pour la reconnaissance de rire et de sourire pour les systèmes statistiques.

Synthèse des systèmes automatiques

Dans cette dernière partie de ma thèse, les systèmes automatiques ont été conçus pour la détection de la rotation de la tête, la détection de l'attention et la détection de rire et sourire sur le corpus ROMEO2 des personnes âgées en interaction avec un robot.

Nous avons d'abord présenté une technique de détection de la rotation de la tête sur la base de la définition d'un ensemble de règles sur les sorties de trois détecteurs de visage de différentes orientations. Le système est évalué sur un sous-ensemble du corpus ROMEO2 et le corpus standard Pointing04. Pour mieux comprendre les raisons pour lesquelles un sujet tourne la tête pendant la conversation avec un robot, nous essayons également d'établir un lien entre la rotation de la tête et des dimensions sociales. Les expérimentations montrent que la rotation de la tête peut être efficacement utilisée pour détecter la perte de l'attention du sujet dans l'interaction avec le robot.

Le deuxième chapitre de cette partie de ma thèse présente un système de détection audio-visuelle de l'attention. L'objectif est de permettre au robot de détecter si le sujet est attentif au robot dans l'interaction et d'adapter son comportement au sujet. Nous avons d'abord mis en place un système auditif qui analyse la voix du sujet et détecte si le sujet est réellement en train de parler au robot. La détection de la rotation de la tête est utilisée pour détecter la perte d'attention du sujet. Nous avons ensuite montré le potentiel d'une méthode en cascade qui utilise à la fois les modalités d'une manière complémentaire. Cette méthode donne de meilleurs résultats par rapport au système auditif.

Le dernier chapitre présente un système de détection audio-visuelle de sourire et rire. La fusion directe des systèmes audio et vidéo se produit en deux étapes: d'abord, chaque système de modalité est exécuté séparément pour produire sa décision individuelle sur un segment aligné des signaux audio et vidéo. La décision finale est positive si au moins l'un des deux systèmes délivre une décision positive. Quand les deux systèmes monomodaux utilisent tous le protocole de séparation de données d'apprentissage et de test « leave-one-out », la fusion des deux systèmes améliore significativement la performance d'un système monomodal au niveau de l'évaluation segmentale. Le système de détection audio-visuelle « in the wild » est également étudié dans ce chapitre en appliquant le protocole « leave-one-out » sur les deux modalités. L'amélioration du système de chaque modalité ramènera en effet une amélioration globale de la performance du système de fusion. Les corrélations entre la performance des systèmes et les questionnaires et les annotations pour les 15 personnes âgées testées qui ont au moins 10 évènements de rire et sourire pendant l'expérimentation sont également analysées. Nous avons trouvé une corrélation

négative entre la performance et le nombre d'évènements de rire et sourire pour le système visuel et le système audio-visuel. De plus, les systèmes audio, vidéo et audio-visuel subissent également une baisse de performance quand les personnes âgées sont plus intéressées par l'expérimentation. Ce phénomène peut être expliqué par le fait que les personnes âgées qui sont plus intéressées par l'expérimentation rient plus souvent et sont plus à l'aise donc avec des poses variées. La variété des poses et le manque de données correspondantes amènent des difficultés pour la reconnaissance de rire et de sourire pour les systèmes statistiques.

Conclusion

Dans mon étude pendant la thèse, je me suis concentré particulièrement pour la détection visuelle de l'attention, de rire et sourire pour les personnes âgées en interaction avec un robot, et la fusion audio-visuelle pour la détection.

Nos intérêts des études couvrent alors :

- Extraction des marqueurs affectifs et sociaux dans les interactions humain-robot:
  - Information cognitive – attention
  - Marqueurs affectifs – rire et sourire
- Combinaison d'indices ou de modalités audio et vidéo : fusion précoce ou tardive
- Données réalistes: Corpus ROMEO2 d'interaction avec les personnes âgées (NAO)

Résumé de contribution

Pour comprendre efficacement et modéliser le comportement des personnes très âgées en présence d'un robot, des données pertinentes sont nécessaires. J'ai participé à la collection du corpus ROMEO2 de personnes âgées (27 sujet avec un âge moyen de 85 ans, durée totale 9h) notamment pour l'enregistrement des données visuelles. J'ai conçu un schéma d'annotation avec les collègues et participé à l'annotation, la vérification et la correction des annotations.

Les annotations sont appliquées aux deux flux audio et vidéo. La période expérimentale (éliminant le temps de réponse au questionnaire, de l'interview, etc.) pour ces 25 sujets concerne environ 6,5 heures de données. L'annotation audio contient 1881 tours parole des sujets à NAO et 529 tours de parole des sujets à l'expérimentateur avec une étiquette émotionnelle, l'étiquette émotionnelle est dans l'ordre décroissant de quantité l'émotion neutre, la joie/satisfaction, le doute, la colère, la surprise, la tristesse, etc. L'annotation audio contient également 417 « affect bursts », y compris 210 évènements de rires. L'annotation vidéo consiste en 1162 évènements de rotation de la tête avec différentes orientations, 231 évènements de hochement de la tête dont 106 en horizontal et 125 en vertical, 96 évènements de mouvement du corps, 724 évènement de mouvement de la bouche (y compris 209 rires, 266 sourires avec la bouche ouverte et 98 sourires avec la bouche fermée) et 273 évènements de mouvement de sourcils dont 277 froncements de sourcils et 62 haussement de sourcils.

J'ai effectué une analyse complète des corrélations entre l'âge, la personnalité, l'autonomie et les comportements des personnes âgées à partir des questionnaires

(questionnaire de satisfaction pour l'expérience de la collecte du corpus et questionnaire de satisfaction) et des annotations du corpus ROMEO2. Plusieurs phénomènes liés au comportement des personnes âgées en interaction avec le robot sont trouvés, par exemple :

- Plus les personnes âgées humanisent le robot, plus les personnes sont satisfaites de l'expérimentation. De plus, l'apparence humaine, la précision de la compréhension du robot et le partage de l'empathie ont un effet positif pour faciliter la projection de comportements humains sur le robot
- La synchronisation entre le destinataire de la parole (signe auditif) et la direction de regard (signe gestuel) peut permettre d'utiliser la rotation de la tête comme un indice supplémentaire pour la détection de destinataire de la parole
- Il existe des corrélations positives entre le nombre d'évènements de rire/sourire et l'utilisation de tutoiement, la préférence d'une apparence robot, la considération du robot comme un humain. Nous pouvons conclure que plus un sujet est détendu, plus il exprime de sourires et de rires pendant l'interaction.
- La bouche étirée est souvent utilisée comme un indice d'insatisfaction quand le sujet pense être mal traité.

L'objectif de ma thèse est l'étude de l'interaction affective et sociale entre des personnes âgées et un robot à partir du corpus ROMEO2. La détection de l'attention consiste à percevoir quand le sujet ne s'adresse pas au robot et à adapter le comportement du robot à la situation. Après avoir considéré les difficultés liées aux personnes âgées et les résultats d'analyse obtenus par l'étude des annotations du corpus, nous nous intéressons à la rotation de la tête au niveau de l'indice visuel et à l'énergie et la qualité de voix pour la détection du destinataire de la parole.

J'ai conçu et réalisé une méthode visuelle de détection de la rotation de la tête sur la base de la définition d'un ensemble de règles sur les sorties de trois détecteurs de visage de différentes orientations. Le système est évalué sur un sous-ensemble du corpus ROMEO2 et a obtenu un résultat de 81,5% comme F-mesure et 9,9% pour le ratio d'erreur balancé au niveau de la trame sur la détection de la rotation de la tête, et un taux de reconnaissance segmentale de l'évènement de rotation de la tête de 87,7%. Le système est également évalué sur le corpus standard des orientations de la tête Pointing04, le taux de bonne reconnaissance pour trois classes (gauche, droit, frontal) est de 83,2%.

Pour mieux comprendre les raisons pour lesquelles un sujet tourne la tête pendant la conversation avec un robot, nous essayons d'établir un lien entre la rotation de la tête et de nombreuses dimensions sociales. Par exemple, 39,5% du temps, un sujet a besoin d'aide ou de la confirmation de l'expérimentateur.

La rotation de la tête peut être efficacement utilisée pour détecter la perte de l'attention du sujet dans l'interaction avec le robot. Par contre, quand la tête du sujet fait face au robot, nous ne pouvons pas être sûr que le sujet soit en train de focaliser son attention sur l'interaction, une détection de l'attention avec la modalité auditive sera nécessaire pour accomplir la tâche. Pour la détection de l'attention, le système final fusionne la détection visuelle de la rotation de la tête et la détection auditive de la qualité de voix (conception et réalisation du système audio par mon collègue Mohamed Sehili) en cascade.

Cette méthode est évaluée au niveau de segment sur une partie du corpus ROMEO2. Le taux de bonne reconnaissance moyenne de destinataire de parole atteint 78,80%, ce qui est beaucoup mieux que la performance du système audio seul, en particulier pour détecter si le sujet s'adresse à l'expérimentateur. La rotation de la tête semble être corrélée avec le fait que le sujet ne s'adresse pas au robot.

Au niveau de la détection de rire et sourire, elle peut être utilisée pour l'étude sur le profil du locuteur et leurs émotions. Mes travaux se concentrent sur la détection de rire et sourire dans la modalité visuelle et la fusion des informations des modalités audio-visuelles (conception et réalisation du système audio par mon collègue Mohamed Sehili) afin d'améliorer la performance du système automatique.

Les expressions spontanées sont différentes des expressions actées ou posées à la fois en apparence et en temps de réaction. Cela signifie que les méthodes utilisées pour la reconnaissance des expressions actées ou posées pourraient ne pas fonctionner correctement sur les expressions réalistes. L'évaluation des méthodes sur un corpus réaliste d'interaction sociale entre les personnes âgées et un robot est un défi. Pour la première expérience, un test sur une méthode standard, reconnue et conservative comme l'utilisation de l'extracteur d'indices LBP avec le classifieur SVM du noyau RBF qui est déjà testée sur de nombreux corpus a été mené. La méthode visuelle est évaluée au niveau de frame, du segment et « in the wild » sur une partie du corpus ROMEO2 avec plusieurs protocoles de séparation de données de test et d'entrainement

La fusion directe des systèmes audio et vidéo pour la détection de rire et sourire se produit en deux étapes: d'abord, chaque système de modalité est exécuté séparément pour produire sa décision individuelle sur un segment aligné des signaux audio et vidéo. La décision finale est positive si au moins l'un des deux systèmes délivre une décision positive. Quand les deux systèmes monomodaux utilisent tous le protocole de séparation de données d'apprentissage et de test « leave-one-out », la fusion des deux systèmes améliore significativement la performance d'un système monomodal au niveau de l'évaluation segmentale. Le système de fusion sous protocole « leave-one-out » a obtenu un résultat d'évaluation segmentale de 83,8% comme le rappel, 54,2% comme la précision et 65,9% comme F-mesure. L'évaluation « in the wild »

du système final est également effectuée en appliquant le protocole « leave-one-out » dans la partie vidéo ainsi que la partie audio, toutes les difficultés dues à l'âge des sujets et le corpus d'interaction sociale réaliste augmentent le défi de la détection. La fusion directe des sorties des deux systèmes amène une forte amélioration du rappel et une petite amélioration du ratio d'erreur balancée.

De plus, les corrélations entre la performance des systèmes de détection de rire et de sourire et les questionnaires et les annotations pour les 15 personnes âgées testées sont analysées. Nous avons également trouvé une corrélation négative entre la performance de détection de rire et sourire et le nombre d'évènements de rire et sourire pour le système visuel et le système audio-visuel. Ce phénomène peut être expliqué par le fait que les personnes âgées qui sont plus intéressées par l'expérimentation rient plus souvent et sont plus à l'aise donc avec des poses variées. La variété des poses et le manque de données correspondantes amènent des difficultés pour la reconnaissance de rire et de sourire pour les systèmes statistiques.

Perspectives

Les informations affectives du corpus ROMEO2 sont annotées avec des étiquettes verbales comme la joie, la tristesse, etc. Une nouvelle annotation continue sur les dimensions affectives (valence, activation, dominance) pourra être très intéressante et supporter les études sur la continuité d'état affectif en interaction avec robot et le système automatique de détection correspondant.

Au niveau de l'étude du comportement de la rotation de la tête des personnes âgées dans le corpus ROMEO2, comme nous avons constaté, un évènement de rotation de la tête est souvent accompagné d'une expression d'émotion du sujet, et donc, il serait très intéressant que le robot reconnaisse ces émotions et adapte ses comportements en conséquence. On peut donc conclure que dans une application réaliste d'interaction orale homme-robot, des données multimodales devraient être considérées en raison de la complexité inhérente à ce genre d'interaction.

Au niveau de la détection de l'attention, nous devrions considérer l'utilisation d'un moyen plus efficace pour combiner des signaux sonores et visuels. Par exemple, on peut imaginer l'utilisation de méthodes qui donnent en sortie une probabilité (par exemple un SVM probabiliste) au lieu d'une décision binaire. Nous devrions également annoter et évaluer une plus grande quantité de données et exécuter une évaluation dépendante du sujet pour savoir si elle peut amener une amélioration.

Pour la détection de rire et de sourire, d'autres indices et des extracteurs d'indices de l'état de l'art pourront être envisagés afin d'améliorer la performance de notre système

et de comparer la performance de détection dans les bases de données actées et notre corpus réaliste de personnes âgées en interaction avec un robot. De plus, le changement de stratégies de fusion mérite également d'être investigué.

En dehors de rire et de sourire, la bouche étirée, les hochements (verticaux et horizontaux) de la tête et le froncement de sourcil sont tous les marqueurs affectifs importants dans l'interaction humain-humain et humain-machine. Les corrélations entre la quantité de ces évènements et les réponses aux questionnaires (questionnaire de satisfaction pour l'expérience de la collecte du corpus et questionnaire de satisfaction) sont déjà analysées et montrées dans le sous-chapitre 4.2 « Analyse des annotations ». Le système de détection de ces indices est intéressant à étudier pour mieux comprendre l'état mental des personnes âgées et permet au robot d'adapter son comportement à la situation.

Nous avons étudié les raisons de la rotation de la tête vers l'expérimentateur (la perte de l'attention) des participants pendant l'expérimentation, les différentes raisons sont accompagnées de différentes émotions, par exemple, pour demander une répétition de la phrase du robot, la parole du sujet est probablement avec une émotion doute; et pour partager une émotion, l'émotion exprimée du sujet est souvent très expressive soit d'amusement ou négative de non compréhension. Nous pouvons donc combiner la détection de l'attention avec la détection des émotions et des marqueurs affectifs pour reconnaitre la raison de la perte de l'attention du sujet afin d'adapter la stratégie d'interaction du robot en temps réel. Un premier système intégrant ces différentes modalités en temps réel est en test actuellement au laboratoire.

Publication

[I.] Sehili, M., Yang, F., Leynaert, V., & Devillers, L. (2014). A corpus of social interaction between nao and elderly people. In *5th International Workshop on Emotion, Social Signals, Sentiment & Linked Open Data (ES3LOD2014). LREC*.

[II.] SEHILI, M., Yang, F., & Devillers, L. (2014, November). Attention detection in elderly people-robot spoken interaction. In *Proceedings of the 2014 Workshop on Multimodal, Multi-Party, Real-World Human-Robot Interaction*(pp. 7-12). ACM.

[III.] Yang, F., Sehili, M. A., Barras, C., & Devillers, L. (2015). Smile and Laughter Detection for Elderly People-Robot Interaction. In *Social Robotics* (pp. 694-703). Springer International Publishing.

[IV.] Devillers, L., et al. (2015). Multimodal data collection of humanrobot humorous interactions in the joker project. In *6th International Conference on Affective Computing and Intelligent Interaction (ACII)*.

Référence

[1] Noureddine Aboutabit, Denis Beautemps, Jeanne Clarke, and Laurent Besacier. A hmm recognition of consonant-vowel syllables from lip contours: the cued speech case. In *A HMM recognition of consonant-vowel syllables from lip contours: the Cued Speech case*, page 4, 2007.

[2] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12):2037–2041, 2006.

[3] Timo Ahonen, Esa Rahtu, Ville Ojansivu, and Janne Heikkilä. Recognition of blurred faces using local phase quantization. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[4] Timur R Almaev and Michel F Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 356–361. IEEE, 2013.

[5] Keith Anderson and Peter W McOwan. A real-time automated system for the recognition of human facial expressions. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 36(1):96–105, 2006.

[6] M. Argyle. Methuen, London,, 1998.

[7] Francis R Bach, Gert RG Lanckriet, and Michael I Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, page 6. ACM, 2004.

[8] Themis Balomenos, Amaryllis Raouzaiou, Spiros Ioannou, Athanasios Drosopoulos, Kostas Karpouzis, and Stefanos Kollias. Emotion analysis in man-machine interaction systems. In *Machine learning for multimodal interaction*, pages 318–328. Springer, 2005.

[9] Tanja Bänziger and Klaus R Scherer. Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Blueprint for affective computing: A sourcebook*, pages 271–294, 2010.

[10] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Recognizing facial expression: machine learning and

application to spontaneous behavior. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 568–573. IEEE, 2005.

[11] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Fully automatic facial action recognition in spontaneous behavior. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 223–230. IEEE, 2006.

[12] Sumit Basu, Tanzeem Choudhury, Brian Clarkson, and Alex Pentland. Towards measuring human interactions in conversational settings. In *Proc. IEEE CVPR Workshop on Cues in Communication*, 2001.

[13] Anton Batliner, Christian Hacker, Stefan Steidl, Elmar Nöth, Shona D'Arcy, Martin J Russell, and Michael Wong. " you stupid tin box"-children interacting with the aibo robot: A cross-linguistic emotional speech corpus. In *LREC*, 2004.

[14] Alberto Battocchi, Fabio Pianesi, and Dina Goren-Bar. Dafex: Database of facial expressions. In *Intelligent Technologies for Interactive Entertainment*, pages 303–306. Springer, 2005.

[15] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

[16] Daniel Bernhardt and Peter Robinson. Detecting emotions from connected action sequences. In *Visual Informatics: Bridging Research and Practice*, pages 1–11. Springer, 2009.

[17] Cynthia Breazeal. Robot in society: friend or appliance. In *Proceedings of the 1999 Autonomous Agents Workshop on Emotion-Based Agent Architectures*, pages 18–26, 1999.

[18] Hervé Bredin and Gérard Chollet. Making talking-face authentication robust to deliberate imposture. In *ICASSP*, pages 1693–1696, 2008.

[19] Axel Buendia and Laurence Devillers. From informative cooperative dialogues to long-term social relation with a robot. In *Natural Interaction with Robots, Knowbots and Smartphones*, pages 135–151. Springer, 2014.

[20] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal

information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211. ACM, 2004.

[21] Nick Campbell. Accounting for voice-quality variation. In *Speech Prosody 2004, International Conference*, 2004.

[22] George Caridakis, Lori Malatesta, Loic Kessous, Noam Amir, Amaryllis Raouzaiou, and Kostas Karpouzis. Modeling naturalistic affective states via facial and vocal expressions recognition. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 146–154. ACM, 2006.

[23] Ginevra Castellano, Iolanda Leite, André Pereira, Carlos Martinho, Ana Paiva, and Peter W McOwan. Affect recognition for interactive companions: challenges and design in real world scenarios. *Journal on Multimodal User Interfaces*, 3(1-2):89–98, 2010.

[24] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[25] Ya Chang, Changbo Hu, Rogerio Feris, and Matthew Turk. Manifold based analysis of facial expression. *Image and Vision Computing*, 24(6):605–614, 2006.

[26] Ya Chang, Changbo Hu, and Matthew Turk. Probabilistic expression analysis on manifolds. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–520. IEEE, 2004.

[27] Ya Chang, Marcelo Vieira, Matthew Turk, and Luiz Velho. Automatic 3d facial expression analysis in videos. In *Analysis and Modelling of Faces and Gestures*, pages 293–307. Springer, 2005.

[28] Chloé Clavel, Ioana Vasilescu, Laurence Devillers, Gaël Richard, and Thibaut Ehrette. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 50(6):487–503, 2008.

[29] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S Chen, and Thomas S Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and image understanding*, 91(1):160–187, 2003.

[30] Jeffrey F Cohn. Foundations of human computing: facial expression and emotion. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 233–238. ACM, 2006.

[31] Jeffrey F Cohn, Lawrence Ian Reed, Zara Ambadar, Jing Xiao, and Tsuyoshi Moriyama. Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 1, pages 610–616. IEEE, 2004.

[32] Jeffrey F Cohn and Karen L Schmidt. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 2(02):121–132, 2004.

[33] Timothy F Cootes and Christopher J Taylor. Active shape models 鈥旀€榮 mart snakes 鈥? In *BMVC92*, pages 266–275. Springer, 1992.

[34] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[35] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[36] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, 18(1):32–80, 2001.

[37] Agnes Delaborde and Laurence Devillers. Use of nonverbal speech cues in social interaction between human and robot: emotional and interactional markers. In *Proceedings of the 3rd international workshop on Affective interaction in natural environments*, pages 75–80. ACM, 2010.

[38] Oscar Déniz, M Castrillon, J Lorenzo, L Anton, and Gloria Bueno. Smile detection for user interfaces. In *Advances in Visual Computing*, pages 602–611. Springer, 2008.

[39] Laurence Devillers, Laurence Vidrascu, and Lori Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422, 2005.

[40] Laurence Devillers, Laurence Vidrascu, and Omar Layachi. Automatic detection of emotion from vocal expression. *A Blueprint for an Affectively Competent Agent, Cross-Fertilization Between Emotion Psychology, Affective Neuroscience, and Affective Computing*, pages 232–244, 2010.

[41] Abhinav Dhall, Akshay Asthana, Roland Goecke, and Tom Gedeon. Emotion

recognition using phog and lpq features. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 878–883. IEEE, 2011.

[42] Abhinav Dhall et al. Collecting large, richly annotated facial-expression databases from movies. 2012.

[43] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Karan Sikka, and Tom Gedeon. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 461–466. ACM, 2014.

[44] Gareth J Edwards, Christopher J Taylor, and Timothy F Cootes. Interpreting face images using active appearance models. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 300–305. IEEE, 1998.

[45] P Ekman, WV Friesen, and JC Hager. Facs manual. *A Human Face*, 2002.

[46] Paul Ekman, Wallace V Friesen, Maureen O'Sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712, 1987.

[47] Gabriele Fanelli, Thibaut Weise, Juergen Gall, and Luc Van Gool. Real time head pose estimation from consumer depth cameras. In *Pattern Recognition*, pages 101–110. Springer, 2011.

[48] Mireia Farrús, Javier Hernando, and Pascual Ejarque. Jitter and shimmer measurements for speaker recognition. In *INTERSPEECH*, pages 778–781, 2007.

[49] Beat Fasel and Juergen Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003.

[50] Xiaoyi Feng, Matti Pietikäinen, and Abdenour Hadid. Facial expression recognition based on local binary patterns. *Pattern Recognition and Image Analysis*, 17(4):592–598, 2007.

[51] Hyoun-Joo Go, Keun-Chang Kwak, Dae-Jong Lee, and Myung-Geun Chun. Emotion recognition from the facial image and speech signal. In *SICE 2003 Annual Conference*, volume 3, pages 2890–2895. IEEE, 2003.

[52] Samuel D Gosling, Peter J Rentfrow, and William B Swann. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528, 2003.

[53] Nicolas Gourier, Daniela Hall, and James L Crowley. Estimating face orientation from robust detection of salient facial structures. In *FG Net Workshop on Visual Observation of Deictic Gestures*, pages 1–9. FGnet (IST–2000–26434) Cambridge, UK, 2004.

[54] Hatice Gunes and Massimo Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, volume 4, pages 3437–3443. IEEE, 2005.

[55] Guodong Guo and Charles R Dyer. Learning from examples in the small sample case: face expression recognition. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 35(3):477–488, 2005.

[56] Stefan Hoch, Frank Althoff, Gregor McGlaun, and Gerhard Rigoll. Bimodal fusion of emotional data in an automotive environment. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 2, pages ii–1085. IEEE, 2005.

[57] Kohsia S Huang, Mohan M Trivedi, and Tarak Gandhi. Driver's view and vehicle surround estimation using omnidirectional video stream. In *Intelligent Vehicles Symposium, 2003. Proceedings. IEEE*, pages 444–449. IEEE, 2003.

[58] Spiros V Ioannou, Amaryllis T Raouzaiou, Vasilis A Tzouvaras, Theofilos P Mailis, Kostas C Karpouzis, and Stefanos D Kollias. Emotion recognition through facial expression analysis based on a neurofuzzy network. *Neural Networks*, 18(4):423–435, 2005.

[59] Akinori Ito, Xinyue Wang, Motoyuki Suzuki, and Shozo Makino. Smile and laughter recognition using speech processing and face recognition from conversation video. In *Cyberworlds, 2005. International Conference on*, pages 8–pp. IEEE, 2005.

[60] Varun Jain and James L Crowley. Head pose estimation using multi-scale gaussian derivatives. In *Image Analysis*, pages 319–328. Springer, 2013.

[61] Qiang Ji, Peilin Lan, and Carl Looney. A probabilistic framework for modeling and real-time monitoring human fatigue. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 36(5):862–875, 2006.

[62] Bihan Jiang, Michel F Valstar, and Maja Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Automatic Face &*

Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pages 314–321. IEEE, 2011.

[63] Martin Johansson, Gabriel Skantze, and Joakim Gustafson. Head pose patterns in multiparty human-robot team-building interactions. In *Social Robotics*, pages 351–360. Springer, 2013.

[64] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[65] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46–53. IEEE, 2000.

[66] Ashish Kapoor, Winslow Burleson, and Rosalind W Picard. Automatic prediction of frustration. *International journal of human-computer studies*, 65(8):724–736, 2007.

[67] Kostas Karpouzis, George Caridakis, Loic Kessous, Noam Amir, Amaryllis Raouzaiou, Lori Malatesta, and Stefanos Kollias. Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. In *Artifical intelligence for human computing*, pages 91–112. Springer, 2007.

[68] Michael Katzenmaier, Rainer Stiefelhagen, and Tanja Schultz. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 144–151. ACM, 2004.

[69] Lyndon S Kennedy and Daniel PW Ellis. Laughter detection in meetings. In *NIST ICASSP 2004 Meeting Recognition Workshop, Montreal*, pages 118–121. National Institute of Standards and Technology, 2004.

[70] Michael Kipp. Anvil: The video annotation research tool. *Handbook of Corpus Phonology. Oxford University Press, Oxford (to appear, 2011)*, 2010.

[71] Chris L Kleinke, Armando A Bustos, Frederick B Meeker, and Richard A Staneski. Effects of self-attributed and other-attributed gaze on interpersonal evaluations between males and females. *Journal of experimental social Psychology*, 9(2):154–163, 1973.

[72] Farid Abedan Kondori, Shahrouz Yousefi, Haibo Li, Samuel Sonning, and Sabina Sonning. 3d head pose estimation using the kinect. In *Wireless Communications and Signal Processing (WCSP), 2011 International Conference on*, pages 1–4. IEEE, 2011.

[73] Uwe Kowalik, Terumasa Aoki, and Hiroshi Yasuda. Broaference – a next generation multimedia terminal providing direct feedback on audience 鈥 satisfaction level. In *Human-Computer Interaction-INTERACT 2005*, pages 974–977. Springer, 2005.

[74] Stephen RH Langton, Helen Honeyman, and Emma Tessler. The influence of head contour and nose angle on the perception of eye-gaze direction. *Perception & psychophysics*, 66(5):752–771, 2004.

[75] Jenn-Jier James Lien. *AUTOMATIC RECOGNITION OF FACIAL EXPRESSIONS USING HIDDEN MARKOV MODELS AND ESTIMATION OF EXPRSSION INTENSITY*. PhD thesis, Washington University, St. Louis, 1998.

[76] Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian Fasel, Mark Frank, Javier Movellan, and Marian Bartlett. The computer expression recognition toolbox (cert). In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 298–305. IEEE, 2011.

[77] Gwen C Littlewort, Marian Stewart Bartlett, and Kang Lee. Faces of pain: automated measurement of spontaneousallfacial expressions of genuine and posed pain. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 15–21. ACM, 2007.

[78] Simon Lucey, Ahmed Bilal Ashraf, and Jeffrey F Cohn. *Investigating spontaneous facial action recognition through aam representations of the face*. INTECH Open Access Publisher, 2007.

[79] Michael Lyons, Shota Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 200–205. IEEE, 1998.

[80] Mounira Maazaoui, Karim Abed-Meraim, and Yves Grenier. Blind source separation for robot audition using fixed hrtf beamforming. *EURASIP Journal on Advances in Signal Processing*, 2012(1):1–18, 2012.

[81] Alison MacLeod and Quentin Summerfield. Quantifying the contribution of vision to speech perception in noise. *British journal of audiology*, 21(2):131–141, 1987.

[82] Paul P Maglio, Teenie Matlock, Christopher S Campbell, Shumin Zhai, and Barton A Smith. Gaze and speech in attentive user interfaces. In *Advances in Multimodal Interfaces 鈥 CMI 2000*, pages 1–7. Springer, 2000.

[83] Maurizio Mancini, Laurent Ach, Emeline Bantegnie, Tobias Baur, Nadia Berthouze, Debajyoti Datta, Yu Ding, Stéphane Dupont, Harry J Griffin, Florian Lingenfelser, et al. Laugh when you 欽棡 e winning. In *Innovative and Creative Developments in Multimodal Interaction Systems*, pages 50–79. Springer, 2014.

[84] Benoit Mathieu, Slim Essid, Thomas Fillon, Jacques Prado, and Gaël Richard. Yaafe, an easy to use and efficient audio feature extraction software. In *ISMIR*, pages 441–446, 2010.

[85] Iain McCowan, Daniel Gatica-Perez, Samy Bengio, Guillaume Lathoud, Mark Barnard, and Dong Zhang. Automatic analysis of multimodal group actions in meetings. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(3):305–317, 2005.

[86] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schröder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3(1):5–17, 2012.

[87] Angeliki Metallinou, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Visual emotion recognition using compact facial representations and viseme information. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 2474–2477. IEEE, 2010.

[88] S Moore and R Bowden. Local binary patterns for multi-view facial expression recognition. *Computer Vision and Image Understanding*, 115(4):541–558, 2011.

[89] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):607–626, 2009.

[90] Radosaw Niewiadomski, Jennifer Hofmann, Jérôme Urbain, Tracey Platt, Johannes Wagner, Bilal Piot, Huseyin Cakmak, Sathish Pammi, Tobias Baur, Stephane Dupont, et al. Laugh-aware virtual agent and its impact on user amusement. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 619–626. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

[91] Jean-Marc Odobez and Sileye Ba. A cognitive and unsupervised map adaptation approach to the recognition of the focus of attention from head pose. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 1379–1382. IEEE, 2007.

[92] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

[93] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[94] Jaak Panksepp. *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press, 1998.

[95] Maja Pantic and Ioannis Patras. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 36(2):433–449, 2006.

[96] Maja Pantic and Leon Rothkrantz. Case-based reasoning for user-profiled recognition of emotions from face images. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, volume 1, pages 391–394. IEEE, 2004.

[97] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 5–pp. IEEE, 2005.

[98] Constantine P Papageorgiou, Michael Oren, and Tomaso Poggio. A general framework for object detection. In *Computer vision, 1998. sixth international conference on*, pages 555–562. IEEE, 1998.

[99] Alex Pentland and Tanzeem Choudhury. Face recognition for smart environments. *Computer*, 33(2):50–55, 2000.

[100] Stavros Petridis and Maja Pantic. Audiovisual discrimination between laughter and speech. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 5117–5120. IEEE, 2008.

[101] Stavros Petridis and Maja Pantic. Audiovisual discrimination between speech and laughter: Why and when visual information might help. *Multimedia, IEEE Transactions on*, 13(2):216–234, 2011.

[102] Rosalind W Picard and Roalind Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997.

[103] Robert Plutchik. *Emotion: A psychoevolutionary synthesis*. Harpercollins College Division, 1980.

[104] Donovan Robert and Rossiter John. Store atmosphere: an environmental psychology approach. *Journal of retailing*, 58:34–57, 1982.

[105] James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294, 1977.

[106] Johanna Ruusuvuori. Looking means listening: coordinating displays of engagement in doctor–patient interaction. *Social Science & Medicine*, 52(7):1093–1108, 2001.

[107] Hugues Salamin, Anna Polychroniou, and Alessandro Vinciarelli. Automatic detection of laughter and fillers in spontaneous mobile phone conversations. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, pages 4282–4287. IEEE, 2013.

[108] Klaus R Scherer. Affect bursts. *Emotions: Essays on emotion theory*, pages 161–196, 1994.

[109] Marc Schröder. Experimental study of affect bursts. *Speech communication*, 40(1):99–116, 2003.

[110] Björn Schuller, Florian Eyben, and Gerhard Rigoll. Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech. In *Perception in multimodal dialogue systems*, pages 99–110. Springer, 2008.

[111] Björn Schuller, Stefan Steidl, and Anton Batliner. The interspeech 2009 emotion challenge. In *INTERSPEECH*, volume 2009, pages 312–315, 2009.

[112] Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456. ACM, 2012.

[113] Jean-Luc Schwartz, Pierre Escudier, and Pascal Teissier. Multimodal speech: Two or three senses are better than one. *Spoken Language Processing*, pages 377–415.

[114] Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. Emotion recognition based on joint visual and audio cues. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 1, pages 1136–1139. IEEE, 2006.

[115] Nicu Sebe, Michael S Lew, Yafei Sun, Ira Cohen, Theo Gevers, and Thomas S Huang. Authentic facial expression analysis. *Image and Vision Computing*, 25(12):1856–1863, 2007.

[116] Mohamed Sehili, Fan Yang, V Leynaert, and L Devillers. A corpus of social interaction between nao and elderly people. In *5th International Workshop on Emotion, Social Signals, Sentiment & Linked Open Data (ES3LOD2014)*. LREC, 2014.

[117] Mohamed El Amine SEHILI, Fan Yang, and Laurence Devillers. Attention detection in elderly people-robot spoken interaction. In *Proceedings of the 2014 Workshop on Multimodal, Multi-Party, Real-World Human-Robot Interaction*, pages 7–12. ACM, 2014.

[118] Thibaud Senechal, Vincent Rapp, Hanan Salam, Renaud Seguier, Kevin Bailly, and Lionel Prevost. Combining aam coefficients with lgbp histograms in the multi-kernel svm framework to detect facial action units. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 860–865. IEEE, 2011.

[119] Thibaud Sénéchal, Vincent Rapp, Hanan Salam, Renaud Seguier, Kevin Bailly, and Lionel Prevost. Facial action recognition combining heterogeneous features via multikernel learning. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(4):993–1005, 2012.

[120] Yusuke Shinohara and Nobuyuki Otsu. Facial expression recognition using fisher weight maps. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 499–504. IEEE, 2004.

[121] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 46–51. IEEE, 2002.

[122] Gabriel Skantze and Joakim Gustafson. Attention and interaction control in a human-human-computer dialogue setting. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 310–313. Association for Computational Linguistics, 2009.

[123] Mingli Song, Jiajun Bu, Chun Chen, and Nan Li. Audio-visual based emotion recognition-a new approach. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–1020. IEEE, 2004.

[124] You Song, Ye Zhang, Zeqiang Wang, and Pengzhi Xie. The head-trace mouse for elderly: A human-computer interaction system based on detection of poses of head and mouth. *International Journal of Information Technology*, 19(2), 2013.

[125] Rainer Stiefelhagen. Tracking focus of attention in meetings. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, page 273. IEEE Computer Society, 2002.

[126] Marie Tahon, Agnès Delaborde, Claude Barras, and Laurence Devillers. A corpus for identification of speakers and their emotions. *LREC, Valetta, Malta*, 2010.

[127] Marie Tahon and Laurence Devillers. Acoustic measures characterizing anger across corpora collected in artificial or natural context. In *Proceedings of the Fifth International Conference on Speech Prosody*, 2010.

[128] Marie Tahon, Mohamed A Sehili, and Laurence Devillers. Cross-corpus experiments on laughter and emotion detection in hri with elderly people. In *Social Robotics*, pages 633–642. Springer, 2015.

[129] James W Tankard Jr. Effects of eye position on person perception. *Perceptual and motor skills*, 31(3):883–893, 1970.

[130] Hai Tao and Thomas S Huang. Explanation-based facial motion tracking using a piecewise bezier volume deformation model. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 1. IEEE, 1999.

[131] Ying-li Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):97–115, 2001.

[132] Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *The journal of machine learning research*, 1:211–244, 2001.

[133] Mohan M Trivedi. Human movement capture and analysis in intelligent environments. *Machine Vision and Applications*, 14(4):215–217, 2003.

[134] Mohan Manubhai Trivedi, Kohsia Samuel Huang, and Ivana Mikic. Dynamic context capture and distributed video arrays for intelligent spaces. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 35(1):145–163, 2005.

[135] K Truong and D Van Leeuwen. Evaluating automatic laughter segmentation in meetings using acoustic and acoustics-phonetic features. In *Proc Proc Workshop on the Phonetics of Laughter at the 16th International Congress of Phonetic Sciences (ICPhS)*, pages 49–53, 2007.

[136] Khiet P Truong and David A Van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144–158, 2007.

[137] Jilin Tu, Thomas Huang, and Hai Tao. Accurate head pose tracking in low resolution video. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 573–578. IEEE, 2006.

[138] http://mplab.ucsd.edu. The MPLab GENKI Database, GENKI-4K Subset.

[139] Michel Valstar and Maja Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65, 2010.

[140] Michel Valstar, Maja Pantic, and Ioannis Patras. Motion history for facial action detection in video. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 1, pages 635–640. IEEE, 2004.

[141] Michel F Valstar, Bihan Jiang, Marc Mehu, Maja Pantic, and Klaus Scherer. The first facial expression recognition and analysis challenge. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 921–926. IEEE, 2011.

[142] Michel F Valstar, Maja Pantic, Zara Ambadar, and Jeffrey F Cohn. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 162–170. ACM, 2006.

[143] Stephanie HM Van Goozen, Nanne E Van de Poll, and Joseph A Sergeant. *Emotions: Essays on emotion theory*. Psychology Press, 2014.

[144] Roel Vertegaal, Robert Slagter, Gerrit Van der Veer, and Anton Nijholt. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 301–308. ACM, 2001.

[145] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.

[146] Michael Voit, Kai Nickel, and Rainer Stiefelhagen. A bayesian approach for multi-view head pose estimation. In *Multisensor Fusion and Integration for Intelligent Systems, 2006 IEEE International Conference on*, pages 31–34. IEEE, 2006.

[147] Alex Waibel, Tanja Schultz, Michael Bett, Matthias Denecke, Robert Malkin, Ivica Rogina, Rainer Stiefelhagen, and Jie Yang. Smart: The smart meeting room task at isl. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 4, pages IV–752. IEEE, 2003.

[148] Yongjin Wang and Ling Guan. Recognizing human emotion from audiovisual information. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 2, pages ii–1125. IEEE, 2005.

[149] Joseph Weizenbaum. Eliza 鈥攁 computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[150] Tingfan Wu, Marian S Bartlett, and Javier R Movellan. Facial expression recognition using gabor motion energy filters. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 42–47. IEEE, 2010.

[151] Tingfan Wu, Nicholas J Butko, Paul Ruvolo, Jacob Whitehill, Marian S Bartlett, and Javier R Movellan. Multilayer architectures for facial action unit recognition. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(4):1027–1038, 2012.

[152] Jing Xiao, Tsuyoshi Moriyama, Takeo Kanade, and Jeffrey F Cohn. Robust full-motion recovery of head by dynamic templates and re-registration techniques. *International Journal of Imaging Systems and Technology*, 13(1):85–94, 2003.

[153] Fan Yang, Mohamed A Sehili, Claude Barras, and Laurence Devillers. Smile and laughter detection for elderly people-robot interaction. In *Social Robotics*, pages 694–703. Springer, 2015.

[154] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*, pages 211–216. IEEE, 2006.

[155] Zhihong Zeng, Yun Fu, Glenn I Roisman, Zhen Wen, Yuxiao Hu, and Thomas S Huang. Spontaneous emotional facial expression detection. *Journal of Multimedia*, 1(5):1–8, 2006.

[156] Zhihong Zeng, Yuxiao Hu, Yun Fu, Thomas S Huang, Glenn I Roisman, and

Zhen Wen. Audio-visual emotion recognition in adult attachment interview. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 139–145. ACM, 2006.

[157] Zhihong Zeng, Yuxiao Hu, Ming Liu, Yun Fu, and Thomas S Huang. Training combination strategy of multi-stream fused hidden markov model for audio-visual affect recognition. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 65–68. ACM, 2006.

[158] Zhihong Zeng, Jilin Tu, Ming Liu, Thomas S Huang, Brian Pianfetti, Dan Roth, and Stephen Levinson. Audio-visual affect recognition. *Multimedia, IEEE Transactions on*, 9(2):424–428, 2007.

[159] Zhihong Zeng, Jilin Tu, Brian Pianfetti, Ming Liu, Tong Zhang, Zhenqiu Zhang, Thomas S Huang, and Stephen Levinson. Audio-visual affect recognition through multi-stream fused hmm for hci. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 967–972. IEEE, 2005.

[160] Yongmian Zhang and Qiang Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(5):699–714, 2005.

[161] Zhengyou Zhang. Microsoft kinect sensor and its effect. *MultiMedia, IEEE*, 19(2):4–10, 2012.

[162] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):915–928, 2007.

Titre : Détection de marqueurs affectifs et attentionnels de personnes âgées en interaction avec un robot.

Mots clés : attention, rire et sourire, détection multimodale, personne âgée, corpus réaliste, interaction sociale avec robot

Résumé : Pour comprendre efficacement et modéliser le comportement des personnes très âgées en présence d'un robot, des données pertinentes sont nécessaires. J'ai participé à la collection d'un corpus de personnes âgées notamment pour l'enregistrement des données visuelles. Le système utilisé pour contrôler le robot est un magicien d'Oz, plusieurs scénarios de conversation au quotidien ont été utilisés pour encourager les gens à coopérer avec le robot. Ces scénarios ont été élaborés dans le cadre du projet ROMEO2 avec l'association Approche.

Nous avons décrit tout d'abord le corpus recueilli qui contient 27 sujets de 85 ans en moyenne pour une durée totale de 9 heures, les annotations et nous avons discuté des résultats obtenus à partir de l'analyse des annotations et de deux questionnaires.

Ma recherche se focalise ensuite sur la détection de l'attention et la détection de rire et de sourire. Les motivations pour la détection de l'attention consistent à détecter quand le sujet ne s'adresse pas au robot et à adapter le comportement du robot à la situation. Après avoir considéré les difficultés liées aux personnes âgées et les résultats d'analyse obtenus par l'étude des annotations du corpus, nous nous intéressons à la rotation de la tête au niveau de l'indice visuel et à l'énergie et la qualité de voix pour la détection du destinataire de la parole. La détection de rire et sourire peut être utilisée pour l'étude sur le profil du locuteur et de ses émotions. Mes intérêts se concentrent sur la détection de rire et sourire dans la modalité visuelle et la fusion des informations audio-visuelles afin d'améliorer la performance du système automatique.

Les expressions sont différentes des expressions actées ou posés à la fois en apparence et en temps de réaction. La conception d'un système qui marche sur les données réalistes des personnes âgées est encore plus difficile à cause de plusieurs difficultés à envisager telles que le manque de données pour l'entrainement du modèle statistique, l'influence de la texture faciale et de la façon de sourire pour la détection visuelle, l'influence de la qualité vocale pour la détection auditive, la variété du temps de réaction, le niveau de compréhension auditive, la perte de la vue des personnes âgées, etc.

Les systèmes de détection de la rotation de la tête, de la détection de l'attention et de la détection de rire et sourire sont évalués sur le corpus ROMEO2 et partiellement évalués (détections visuelles) sur les corpus standard Pointing04 et GENKI-4K pour comparer avec les scores des méthodes de l'état de l'art.

Nous avons également trouvé une corrélation négative entre la performance de détection de rire et sourire et le nombre d'évènement de rire et sourire pour le système visuel et le système audio-visuel. Ce phénomène peut être expliqué par le fait que les personnes âgées qui sont plus intéressées par l'expérimentation rient plus souvent et sont plus à l'aise donc avec des poses variées. La variété des poses et le manque de données correspondantes amènent des difficultés pour la reconnaissance de rire et de sourire pour les systèmes statistiques.

Les expérimentations montrent que la rotation de la tête peut être efficacement utilisée pour détecter la perte de l'attention du sujet dans l'interaction avec le robot. Au niveau de la détection de l'attention, le potentiel d'une méthode en cascade qui utilise les modalités d'une manière complémentaire est montré. Cette méthode donne de meilleurs résultats que le système auditif seul. Pour la détection de rire et sourire, en suivant le même protocole « Leave-one-out », la fusion des deux systèmes monomodaux améliore aussi significativement la performance par rapport à un système monomodal au niveau de l'évaluation segmentale.

Title : audio-visual detection of emotional (laugh and smile) and attentional markers for elderly people in social interaction with a robot.

Keywords : attention, laughter and smile, multimodal detection, elderly people, realistic corpus, social interaction with robot

Abstract : To effectively understand and model the pattern of behavior of very old people in the presence of a robot, relevant data are needed. I participated in the collection of a corpus of elderly people in particular for recording visual data. The system used to control the robot is a Wizard of Oz, several daily conversation scenarios were used to encourage people to interact with the robot. These scenarios were developed as part of the ROMEO2 project with the Approche association.
We described at first the corpus collected which contains 27 subjects of 85 years' old on average for a total of 9 hours, annotations and we discussed the results obtained from the analysis of annotations and two questionnaires.
My research then focuses on the attention detection and the laughter and smile detection. The motivations for the attention detection are to detect when the subject is not addressing to the robot and adjust the robot's behavior to the situation. After considering the difficulties related to the elderly people and the analytical results obtained by the study of the corpus annotations, we focus on the rotation of the head at the visual index and energy and quality vote for the detection of the speech recipient. The laughter and smile detection can be used to study on the profile of the speaker and her emotions. My interests focus on laughter and smile detection in the visual modality and the fusion of audio-visual information to improve the performance of the automatic system.
Spontaneous expressions are different from posed or acted expression in both appearance and timing. Designing a system that works on realistic data of the elderly is even more difficult because of several difficulties to consider such as the lack data for training the statistical model, the influence of the facial texture and the smiling pattern for visual detection, the influence of voice quality for auditory detection, the variety of reaction time, the level of listening comprehension, loss of sight for elderly people, etc.
The systems of head-turning detection, attention detection and laughter and smile detection are evaluated on ROMEO2 corpus and partially evaluated (visual detections) on standard corpus Pointing04 and GENKI-4K to compare with the scores of the methods on the state of the art.
We also found a negative correlation between laughter and smile detection performance and the number of laughter and smile events for the visual detection system and the audio-visual system. This phenomenon can be explained by the fact that elderly people who are more interested in experimentation laugh more often and therefore perform more various poses. The variety of poses and the lack of corresponding data bring difficulties for the laughter and smile recognition for our statistical systems.
The experiments show that the head-turning can be effectively used to detect the loss of the subject's attention in the interaction with the robot. For the attention detection, the potential of a cascade method using both methods in a complementary manner is shown. This method gives better results than the audio system. For the laughter and smile detection, under the same leave-one-out protocol, the fusion of the two monomodal systems significantly improves the performance of the system at the segmental evaluation.