Contextualisation de descripteurs affectifs à partir de scénarios d'interaction avec un assistant virtuel

Romain Fernandez

Contextualisation de descripteurs affectifs à partir de scénarios d'interaction avec un assistant virtuel

Romain FERNANDEZ

Sous la direction de Fabien Ringeval

Laboratoire : LIG

UFR LLASIC
Département Sciences du Langage

Mémoire de master 2 Science du Langage – 20 crédits
Parcours : Industrie de la Langue, orientation recherche
Année universitaire 2022-2023

Contextualisation de descripteurs affectifs à partir de scénarios d'interaction avec un assistant virtuel

Romain FERNANDEZ

Sous la direction de Fabien Ringeval

Laboratoire : LIG

UFR LLASIC
Département Sciences du Langage

Mémoire de master 2 Science du Langage – 20 crédits
Parcours : Industrie de la Langue, orientation recherche
Année universitaire 2022-2023

Remerciements

Je tiens à remercier en premier lieu Fabien Ringeval, qui a accepté de m'encadrer et de m'aider lors de mon stage. Il m'a beaucoup appris pendant cette expérience et j'en ressors grandi.

Je voulais également remercier Sina, Hippolyte, Yongxin et Safa de m'avoir accueilli au sein de l'équipe THERADIA, de s'être montrés bienveillants et patients, et de m'avoir conseillé et aidé tout au long de mon stage.

Je remercie également le LIG pour l'accueil, les conseils et l'ambiance détendue au sein du laboratoire.

Je remercie enfin ma famille et mes amis de m'avoir soutenu et encouragé tout au long de l'année.

Enfin, un remerciement spécial à Fanny, sans qui cette année aurait été beaucoup plus difficile à gérer qu'elle ne l'était déjà.

DÉCLARATION ANTI-PLAGIAT

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

PRENOM : Romain

NOM :Fernandez

DATE : 25 / 07 / 2023

Table des matières

INTRODUCTION

L'objet de ce mémoire vise à explorer l'apport du contexte dans le cadre d'interaction en dyade avec un assistant virtuel. C'est un champ particulièrement jeune et fécond des recherches en *Affective Computing.* Cette recherche s'est effectuée dans le cadre du projet THERADIA. Projet financé par la BPI (Projet Structurant pour la Compétitivité) sur 5 ans visant à développer un assistant virtuel accompagnant des sujets atteints de troubles cognitifs ou de maladies neurodégénératives (Alzheimer par exemple) lors de séances de thérapie cognitive. Les données récoltées par l'assistant virtuel lors de ces séances de thérapie sont partagées avec les aidants et les médecins spécialistes afin d'assurer un meilleur suivi des patients.

Chaque donnée récoltée dans le cadre de ce projet a été annotée afin de faire partie d'une taxonomie créée par Maud Costa (vacataire, SdL UGA) décrivant les différents scénarios d'interaction apparaissant au cours d'une thérapie. À l'aide de descripteurs affectifs (ce sont ici les FAUs qui seront utilisées ; à la fois car l'on émet l'hypothèse qu'elles contiennent énormément d'informations, mais également parce que nous manquons de temps pour exploiter d'autres descripteurs multi-modaux), nous avons entraîné des SVM afin de reconnaître automatiquement un type de contexte particulier. Si les affects sont influencés par le contexte, alors on pourrait s'attendre à ce que les expressions faciales dépendent également d'un contexte et de stimuli extérieurs. Les modèles entraînés nous apparaissent comme un outil nous permettant d'évaluer la reconnaissance d'un contexte en particulier, mais également à l'avenir la détection de contexte nécessitant l'intervention de l'assistant virtuel développé dans le cadre du projet THERADIA.

Ce mémoire est composé de trois parties principales : en premier lieu l'état de l'art afin de rentre compte de l'état de la recherche sur le contexte en sciences sociales et en affective computing. Dans un second temps, le protocole de récolte de données et de création

de corpus sera évoqué, et enfin nous aborderons la modélisation des scénarios d'interaction via les SVM en partant de statistiques préliminaires pour aboutir aux résultats des différents modèles.

ETAT DE L'ART

1. CADRE THÉORIQUE

Dans cette première partie de l'état de l'art nous nous attacherons à faire un tour d'horizon des différentes avancées scientifiques à travers le temps sur des notions clefs autour de l'interaction humain-machine.

1.1. Concept d'interaction et son informatisation

Les sciences humaines, soit la psychologie, la psychologie sociale ou la sociologie, ont cherché à décrire les mécaniques sociales et cognitives derrières les interactions en dyade ou en groupe, mais aussi les émotions et les signaux sociaux. C'est grâce aux avancées dans ces domaines que l'on a pu mettre au point les premiers systèmes d'interaction humain-machine.

1.1.1. De l'interaction sociale Humain-Humain

La psychologie sociale a longtemps cherché à schématiser, formaliser et structuraliser le concept d'interaction. À la jonction entre le XIX^{e} et le XX^{e}, l'interaction sociale s'effaçait au profit d'un autre concept : l'action sociale. Concept théorisé par Max Weber (*Économie et société, 1978*)[1]. Cette action sociale est l'émission d'un comportement envers un autrui animé ou non. Par ce fait, Weber entendait comprendre les faits sociaux par une simplification des interactions entre les individus. L'interaction est alors considérée dans sa forme macro-structurelle et non en ce qu'elle contient à l'échelle micro-structurelle, c'est à dire des signaux sociaux. Au milieu du XX^{e} siècle, Mead propose une nouvelle conception de l'interaction qu'il nomme en certain cas « triadic matrix ». (G.H Mead, cité par J.H Turner, 1988 : 74)[2]. Il faut comprendre dans l'idée de cette matrice triadique qu'il n'y a que deux entités : une première qui se meut dans un environnement et ainsi envoie des signaux à d'autres entités, et une deuxième qui perçoit ces mouvements, les segmente en gestes et y répond en modifiant ses propres mouvements, qui deviennent à leur tour des signaux envoyés à la première entité. Cette dernière recevra ces signaux et modifiera son comportement et ainsi de suite. Ce qu'apporte Mead avec ce concept, c'est déjà une structuration de l'interaction non pas à l'échelle humaine mais s'appliquant à toutes les interactions y compris animales. Ce schéma

permet également de comprendre les échanges interactionnels entre les individus et la nécessité de recevoir et d'interpréter des signaux afin d'adapter son comportement.

C'est sur ce modèle prototypique que la psychologie sociale s'attardera sur les composantes de l'interaction sociale entre les humains et développera de nouveaux modèles d'interaction mais également d'émotions. Une des avancées les plus marquantes de la psychologie sociale dans le champ scientifique à l'heure actuelle est le concept d'*Engagement*, introduit sous un autre nom en 1987 par Coker D. et Burgoon J. dans leur ouvrage *The Nature of conversational involvement and non-verbal encoding patterns*[3]. Oertel et al. (2022)[4] considèrent que cette notion de *conversational involvement* est identique à celle d'engagement dans leur article, cette dernière étant relativement moderne. Pour la première fois, l'interaction ne se limite pas à l'échange et la modification de comportements gestuels et verbaux. Il s'agit également de créer une connexion avec l'autre puis de la maintenir pendant la durée de l'échange. La notion d'engagement englobe et fait partie intégrante des schémas d'interactions actuels. S'il n'y a ni traduction ni définition claire et précise de l'engagement on lui trouve néanmoins plusieurs synonymes servant de définition contextuelle comme : « l'intérêt, l'attention soutenue, immersion et l'implication ». (Oertel et al. 2022 : 2, traduction personnelle). L'on doit cette apparition fulgurante de l'engagement dans le champ de la recherche aux nombreux travaux effectués sur les signaux sociaux verbaux et non-verbaux envoyés entre les individus pour communiquer depuis le XX^{e} siècle. Les nombreuses terminologies et théories sémantiques qui fleurissent sur ces signaux ont permis d'affiner l'analyse des mécanismes relationnels. Argyle, dans son ouvrage *Psychology of interpersonal behaviour*[5] affirme l'importance des comportements gestuels dans le cadre de relations sociales. De la même manière Ekman et Friesen[6] (Ekman et Friesen cités par Pantic et al.[7], 2008, p : 171-172, traduction personnelle.) proposent une terminologie des comportements non-verbaux et des intentions communicationnelles qu'ils véhiculent :

- « Les états affectifs, attitudinales et/ou cognitifs (peur, joie, stress, désagrément, etc.)
- Les emblèmes (les signaux interactionnels d'une culture, comme le pouce levé.)
- Manipulateurs (actions qui servent à manipuler, influencer un objet dans l'environnement, ou des actions d'auto-manipulations comme se mordre les lèvres, se gratter)

- Illustrateurs (actions qui accompagnent le discours comme pointer du doigt)
- régulateurs (médiateurs de conversations comme hocher la tête, sourire etc.) »

Ces intentions sont les paramètres qui modifient les comportements des individus dans le cadre d'une relation. Dans la schéma de Mead cité plus-haut la seconde entité recevant des signaux décode ces intentions communicationnelles afin d'adapter son comportement. Quant aux signaux régulateurs, ils sont généralement destinés à maintenir l'*Engagement* de l'interlocuteur. Ces signaux peuvent être de plusieurs natures et ont été de nombreuses fois étudiés dans la littérature psychologique car ils sont à l'origine de toutes interactions.

1.1.2 … à l'interaction sociale Humain-Machine

Depuis le milieu du XX^{e} siècle, plusieurs tentatives d'agent conversationnel sont apparues, de la très célèbre Éliza dans les années 60 au chat-bot modernes sur les sites web. L'émergence de ces agents a suscité beaucoup de questions théoriques et pratiques sur l'interaction entre l'humain et la machine. Si au départ cette interaction se faisait par le prisme de l'ordinateur, aujourd'hui elle peut se faire par des canaux multimodaux comme la vidéo et l'audio. Mahmud et al. [17] séparent ces interactions selon deux types d'interface : l'interface tangible (le clavier et la souris, comme pour interagir avec Éliza) et l'interface intangible (basée sur la vision, les expressions faciales etc.) Dans le premier cas, l'interaction se fait uniquement par le texte et sur une analyse de surface, dans le second, l'agent conversationnel prend en compte différents types de signaux sociaux pour adapter son comportement. Les efforts de la recherche dans les HMI (Interaction Humain-Machine) se concentrent davantage sur les interfaces intangibles qui correspondent à des systèmes intelligents interactionnels centrés sur l'homme (HCI)^{2}[7].

Ainsi il existe une terminologie des agents virtuels qui définissent leurs rôles et fonctions[12]. L'agent virtuel peut simplement fournir des informations, comme présenter la météo, sans qu'il y ait de réelle interaction. L'agent peut également dialoguer avec un utilisateur comme un chat-bot sur un site web. Dialogue qui se fait soit par bulles de dialogue soit par canaux multimodaux. On peut simuler un jeu de rôle (dans un environnement entièrement virtuel) et des types de conversation : l'agent serait un vendeur et l'utilisateur un

acheteur. La discussion est limitée par son contexte mais permet davantage de précision dans l'interaction. Enfin on peut simuler des conversations de groupe avec un ou plusieurs agents conversationnels. Ces divisions sont à titres indicatives : un chat bot peut être dans un rôle d'enseignant ou bien faire passer des entretiens d'embauche aux utilisateurs comme c'est le cas de Susanne [18], un agent conversationnel virtuel doué de talents en communication verbale et non-verbale et qui permet d'améliorer les capacités sociales d'utilisateurs en les mettant en situation d'embauche. Dans leur article sur la modélisation et l'analyse des comportements humains dans le cadre d'interactions HMI, Vinciarelli et al.[19 : 7] cite une étude du NISR-TG (*Natural Interaction with Social Robots Topic Group*) qui décrit les différents niveaux d'interactions des agents conversationnels et des robots :

Tableau 1: Taxonomie des niveaux d'interactions des agents conversationnels

Nv. 0	L'agent n'interagit pas avec l'humain
Nv. 1	L'agent perçoit l'humain comme un objet
Nv. 2	L'agent perçoit l'humain comme un autre agent représenté explicitement, et peut-être ré-identifié dans le temps
Nv. 3	Interaction dans les deux sens mais l'humain doit connaître et obéir à des conventions et des comportements reconnus par le système de l'agent
Nv. 3a	Interaction à double sens possible avec la capacité de parler
Nv. 4	L'agent adapte son comportement à celui du partenaire pendant l'interaction
Nv. 5	L'agent reconnaît différents utilisateurs et en fonction adapte son comportement
Nv. 6	L'agent est capable d'interagir avec plus d'un utilisateur
Nv. 7	L'agent à des traits de personnalités que l'utilisateur peut reconnaître. Induit différents comportements dans des situations identiques
Nv. 8	L'agent est capable d'apprendre et d'accumuler de l'expérience à travers plusieurs interactions
Nv. 9	L'agent est capable de construire et de maintenir une relation avec l'utilisateur

Les recherches sur le domaine des émotions, du traitement des signaux sociaux et de l'affective computing en général tendent à modéliser et créer des agents conversationnels qui correspondent au minima à un niveau 4 c'est à dire des agents conversationnels capables

d'interagir et d'interpréter les informations que leurs envoie l'utilisateur à travers sa posture, son regard, ses mains et son discours.

1.2. Des signaux Sociaux au SSP

L'on peut répartir les signaux sociaux en deux grandes catégories : d'une part les signaux verbaux qui comprennent l'énoncé d'un locuteur mais également la manière de le dire et d'autre part les signaux non-verbaux qui se résument à des comportements gestuels. Ces deux catégories ne sont pas exclusives, bien au contraire. Dans le cas d'interactions sociales elles sont souvent complémentaires et permettent de mieux transmettre une information. L'arrivée de l'informatique dans les sciences sociales a permis la modélisation, l'explicitation et l'amélioration des modèles théoriques en psychologie et sociologie. Rendre tangible ces dits modèles permettait au-delà de l'application pratique de remettre en cause le paradigmes existants en explicitant certains *a priori* dans des modèles émotionnels ou en mesurant et en exploitant les signaux sociaux. Dès lors une grande complicité s'est créée entre les deux domaines de recherche, s'alimentant mutuellement afin d'atteindre des modèles de plus en plus complexes mais également de plus en plus concrets.

1.2.1. Les Signaux Verbaux

Concernant les signaux verbaux, deux sous-catégories émergent [7]. D'une part les signaux linguistiques (explicites) qui contiennent l'ensemble des informations contenues dans une langue à sa surface. À ce sujet, Pantic et al. [7] citent le travail de Whissel effectué en 1989, le *Dictionnary of Affect Language*. Ce dictionnaire contient plusieurs milliers de mots dont la valeur affective a été évaluée puis testée sur un ensemble d'expériences afin d'assurer la fiabilité de cette notation. Cependant, Pantic et al. rappellent que l'on ne peut compter sur cette aspect purement de surface afin de prédire le choix de mots d'une personne et le comportement affectif associé. D'autre part, les signaux paralinguistiques (implicites) ont montré une plus grande efficacité quant à la prédiction d'affects ou d'émotions. Le texte de Justin P.N et Scherer K.R.[8] cité par Pantic et al. démontre qu'en général des interlocuteurs extraient et décodent précisément des informations sur l'état affectif d'un locuteur comme l'ennui via sa prosodie. On retrouve parmi les signaux paralinguistiques plusieurs indices

permettant de décrypter l'intention communicationnelle comme le pitch, la qualité de la voix, l'intensité, etc. [9]

Pour les signaux audios, plusieurs données paralinguistiques peuvent également être extraites de l'onde sonore. Valstar et al.[21] utilisent l'outil COVAREP pour extraire différentes informations. Ainsi ils obtiennent le pitch (la fréquence fondamentale), le voisement, diverses mesures sur la qualité de la voix (différences en amplitude des deux premières harmoniques des différents spectres glotaux H1 H2, la variation spectrale ou *peak-slope*…), des coefficients cepstraux etc. Vinciarelli et al. [22] proposent dans leur étude sur le traitement des signaux sociaux d'ajouter comme comportement non-vocal majeur les silences (d'hésitation, psycholinguistique : quand le locuteur cherche ses mots, ou d'interaction) ainsi que les tours de paroles.

1.2.2. Les Signaux Non-Verbaux

Les signaux non-verbaux sont étudiés et fortement documentés depuis le milieu du XX^{e} siècle. Si ils concentrent tant d'effort, c'est notamment car longtemps ignorés, ils révèlent en réalité une grande quantité d'informations utiles à la communication mais également à la compréhension de l'interaction [14]. Cette catégorie de signaux sociaux peut également être subdivisée en trois grandes familles : l'étude des expressions faciales, des mains et de la posture corporelle.

- Les Expressions faciales

Ekman & Friesen propose dès 1978 le FACS [10], le *Facial Action Coding System*. Ce FACS décompose le visage en plusieurs zones d'actions, des parties du visage qui s'activent lorsque l'on essaie de communiquer une information quelconque. Ce système sera augmenté par la suite sous le nom d'EMFACS qui permettra de réduire le nombre d'unités d'action (soit les zones d'activation) et de les associer à des émotions particulières. Bien qu'il soit remis en cause, ce système sera sollicité à de nombreuses reprises dans les études portant sur les interactions sociales [11] en psychologie mais également pour la détection automatique d'émotion sur des données vidéos.

Comme pour les signaux audios, les signaux vidéos peuvent être traités par différents outils. Dans l'AVEC 2016, Valstar et al.[21] utilisent openFace pour reconnaître des parties du

visage en activation et les segmenter. À partir de là ils obtiennent différentes mesures, comme l'approximation du regard pour les deux yeux, la posture de la tête en 3D et son orientation, ainsi que les différentes AU (*Actions Units*) du FACS repérées par le software FACET qui associe à ces AU des émotions. Les données vidéos sont également traitées dans leurs dynamiques[22] au moyen d'HMM, de champs aléatoires conditionnelles ou de la déformation temporelle dynamique, afin de pouvoir estimer la rapidité du geste et sa fluidité. Ces critères permettront par la suite de classifier avec davantage de rigueur les affects, les émotions et les intentions de l'utilisateur.

- Les Gestes

En ce qui concerne les gestes, plusieurs taxonomie sont en vigueur. On peut citer celle de McNeill [13] considérée comme une des plus célèbres selon l'article d'André E. et Pelachaud C. [12 : 8] :

> « He[McNeil] defined several communicative gesture types: iconic (refers to some physical/spatial/temporal properties of the speech referents), metaphoric (refers to abstract property of the speech referents), deictic (indicates a concrete or abstract point in the world), and beat (follows the rhythmic structure of the speech), and emblem (has well-specific meaning) »

Cette taxonomie permet non seulement de décrire mais également d'annoter sémantiquement les différents gestes effectués avec les mains. Cela peut être des marqueurs d'emphase, insistant sur un élément précis de l'énoncé, ou bien un pattern émotif ou affectif (comme peut l'être l'emblème : taper ou serrer les poings).

- Posture physique et Physiologie

Enfin la posture physique tend également à véhiculer des informations sur le comportement ou l'intention d'une personne. Ces postures sont signifiantes [15] car elles renseignent sur l'ordre de l'interaction, les tours de paroles dans des groupes, mais aussi sur le lien et l'*engagement* entretenus entre un locuteur et un interlocuteur. Ces analyses de postures sont facilitées avec l'arrivée des nouvelles technologies qui permettent de simplifier et modéliser les squelettes des postures afin d'en extraire davantage d'informations y compris dans leurs dynamiques.

Il faudrait rajouter à cette liste les nombreux indices physiologiques que l'on obtient via des capteurs spéciaux comme le rythme cardiaque, le rythme respiratoire, la transpiration,

etc. qui permettent d'inférer des hypothèses sur l'état mental d'un personne à un instant *t*. Ces indices sont de plus en plus utilisés dans la littérature scientifique, en psychologie mais également en informatique lors d'expériences.

1.2.3. Architecture exploitant les signaux sociaux

Ces données sont utilisées par les modèles pour pouvoir classifier des émotions, reconnaître des traits de personnalités. Le SSI Framework [20] est une architecture *Blackboard* qui se compose d'un *Stream* (l'ensemble des données capturés par les capteurs), qu'il décompose en *Events* (ce sont les segments qui seront utilisés pour la classification). Ces segments sont stockés dans une *Event Board* (base de données des différents segments qui nourrit par la suite la classification). Néanmoins le SSI Framework n'apprend pas à prendre en considération le contexte nécessaire lors de la création d'un segment.

C'est à cette question que répondent les modèles utilisant le Deep-Learning comme le modèle MARSSI, fondé sur une architecture RNN avec des cellules LSTM qui permettent d'ajuster la durée des segments en fonction du contexte nécessaire. Ce modèle donne la possibilité à l'agent virtuel de se montrer plus empathique et de réfléchir aux différentes stratégies de régulation en prenant en compte des comportements passés pour actualiser l'analyse d'un comportement présent. Les nombreuses recherches faîtes dans le domaines du Deep-Learning permettent d'accroître les capacités d'identification, de classification et d'adaptation des agents conversationnels en temps réel [24].

2. CADRE MÉTHODOLOGIQUE

Cette partie sera consacrée au recueil de données, leur traitement ainsi qu'à l'évaluation des systèmes de prédictions automatiques.

2.1. Données

Les données sont essentielles pour nourrir des modèles de reconnaissance et de génération automatique de signaux sociaux. Elles sont très souvent à l'origine des performances d'un système, ce qui les rend non-négligeables.

2.1.1. Création de corpus

Malgré le fort enthousiasme de la recherche à proposer et améliorer des systèmes de reconnaissance automatique de la parole et de la vidéo, il y a un manque crucial de données. De fait les chercheurs ont à plusieurs reprises créé des corpus afin de répondre aux exigences des modèles envisagés. Lors du challenge ComPaRe 2013[25] quatre corpus ont été utilisés afin de réaliser des modèles prédictifs dans le domaine de l'affective computing :

- *SSPNet Vocalisation Corpus ou SVC*

Contient un ensemble de 2763 clips audio de 11 secondes provenant de 60 appels téléphoniques de 120 participants différents (63 femmes et 57 hommes). L'annotation de ce corpus s'est concentrée sur les rires et les vocalisations de remplissages. Ces vocalisations ont été annotées par un premier auditeur et écoutées par un second qui validait l'annotation ou non.

- *SSPNet conflict corpus ou SC*^{2}[26]

Ce corpus se compose de 1430 clips de 30 secondes extraits eux même du corpus Canal9. Il s'agit de débats politiques suisses en langue française. Le corpus s'intéresse au niveau de conflit dans la communication non-verbale en ne prenant en compte que des locuteurs non-français pour éviter tout biais lors de l'annotation.

- *Geneva Multimodal Emotion Portrayals (GEMEP)*[27]

Le GEMEP est un corpus multimodal dans lequel des acteurs (5 hommes, 5 femmes) jouent des petits scénarios afin d'exprimer une émotion. Le corpus joint un enregistrement à la fois audio et vidéo afin de pouvoir analyser le comportement verbal et non-verbal. De plus, parmi les phrases à prononcer, deux n'avaient aucun sens et devaient simplement être prononcées selon une modalité émotionnelle.

- *Child Pathological Speech Database ou CPSD*[28]

Le CPSD est un corpus regroupant un ensemble de 26 phrases prononcées par 99 enfants entre 6 à 18 ans sur des modalités différentes (déclarative, exclamative etc.) Tous les enfants ont des troubles du développements envahissants. Le corpus ayant été enregistré à différents endroits, les conditions acoustiques selon les données ne sont pas identiques.

Récapitulatif des différentes base de données mentionnées

DataBase	Taille (heure)	Annotation	Contexte
SSPNet Vocalisation Corpus ou SVC	8,4	Vocalisations et rires	Conversations téléphoniques
SSPNet conflict corpus ou SC2	11,5	Tours de paroles, aggressivité, Posture selon un questionnaire	Débats télévisés
Geneva Multimodal Emotion Portrayals (GEMEP)	/	Modalités affectives, Force de l'expression, conviction	Acteurs prononçant des phrases Sur des modalités affectives
Child Pathological Speech Database ou CPSD	1	Intonation et modalité de phrase	Enfant avec troubles Du développement
AMI (Augmented Muti-party Interaction)	100	Transcription, Posture et signaux Non verbaux	Interaction de groupe

Tableau 2: Récapitulatif des bases de données

Comme on peut le constater parmi ces quatre corpus, deux sont multimodaux (le GEMEP et le SC^2). Les deux autres sont purement audio. L'annotation étant un travail long et fastidieux, les modèles s'entraînent souvent sur des corpus libres de droits et qui possèdent un lot de données conséquent comme le SC^2. La recherche sur l'engagement nécessitant des données précises et viables, depuis 2010 plusieurs corpus multimodaux plus petits sont créés. C'est le cas du corpus de Kim J. et al.[29] qui contient 3 heures d'enregistrement d'enfants autistes (entre 5 à 8 ans, effectuant une tâche). Ce corpus a ensuite été utilisé pour détecter automatiquement l'engagement à travers leurs signaux verbaux et non-verbaux. D'autres part, des corpus sont également créés afin de palier le manque de données sur les interactions en groupe. C'est le cas du corpus AMI[30] (*Augmented Muti-party Interaction*) qui contient 100 heures de données audiovisuelles annotées de réunions.

Néanmoins, hormis le cas d'AMI, les corpus créés pour les modèles sont sensiblement petits et ne permettent pas toujours d'optimiser les systèmes de reconnaissance. En ce sens, la démarche du projet THERADIA[31] diffère. En utilisant un magicien d'Oz, plusieurs dizaines de participants ont été enregistrés durant une interaction avec un agent virtuel pendant une à deux heures. Ces sessions ont été par la suite annotées, constituant un corpus conséquent et riche.

2.1.2 Annotations de corpus

L'annotation s'est vue simplifiée avec l'apparition de logiciels tels qu'ELAN ou ANVIL. Ces logiciels ont permis aux chercheurs de segmenter, transcrire, et annoter plus rapidement, simplement et efficacement. À ces logiciels il faut encore ajouter des codes et des

grilles d'annotation qui permettent d'universaliser et de théoriser les domaines que sont l'interaction, l'engagement, la reconnaissance d'émotions, etc.

C'est le cas du MUMIN *Coding Scheme*[32] développé par le *Nordic Network on Multimodal Interfaces*. Il s'agit d'un outil général qui cherche à simplifier l'étude des signaux non-verbaux en particulier les expressions faciales et le mouvement des mains. Le MUMIN *Coding Scheme* permet d'annoter plusieurs caractéristiques de ces signaux : basique (si le signal est perçu, et si oui, s'il est compris par l'interlocuteur), acceptation (si l'interlocuteur accepte ou refuse le signal perçu) et l'émotion (joyeux, triste, dégoût, surpris, colère, effrayé, autre). En plus de décrire ces signaux, MUMIN propose un système de modélisation des tours de paroles précis : on initialise un tour de parole avec *Turn Gain* (qui devient en fonction du contexte *Turn Take* si le locuteur prend la parole de lui-même, ou *Turn Accept* si le locuteur accepte qu'on lui donne la parole). Ce tour de parole est maintenu *Turn Hold* et est terminé par *Turn End* (si le locuteur lâche la parole par pression cela devient *Turn Yield*, s'il offre la parole à quelqu'un d'autre *Turn Offer* et s'il termine son tour *Turn Complete*). Enfin, l'annotation permet de créer des structures en découpant la vidéo en séquence. En plus de ces annotations structurelles s'ajoutent des annotations descriptives sur les mouvements (36 traits d'annotations pour le visage et 7 traits d'annotations pour les mains). Enfin on peut rajouter à ces traits des catégories sémiotiques : *Indexical Deictic, Iconic, Symbolic*, etc. L'objectif de MUMIN est avant tout de fournir un outil généralisant, mais l'on constate qu'il prend en compte le contexte dans les interactions par ses annotations purement structurelles.

D'une autre façon, Saint-Georges et al.[33] contextualisent l'annotation d'une étude pré-existante afin d'améliorer les modèles informatiques développés. Ils ont segmenté et annoté un corpus se composant d'interactions entre des enfants (dont une partie est atteinte de troubles du développement) et leurs parents. Ce corpus avait déjà été utilisé pour étudier les premières années d'enfants diagnostiqués avec ou sans trouble et les interactions qu'ils entretiennent avec leurs parents. Cet article a pour hypothèses que les enfants ayant des troubles du développement comme l'autisme auront un développement social déviant à l'inverse de ceux ayant un handicap mental qui présenteront un retard dans l'apprentissage. Également, l'étude cherche à montrer que les parents des enfants ayant des troubles du développement constatent très tôt les soucis de développement de leurs enfants et induisent chez eux des patterns d'interactions particuliers. Comme le souligne les chercheurs, l'étude préliminaire n'avait pas pris en considération le contexte de l'interaction (le comportement

des parents, le comportement des enfants, le comportement des parents envers les enfants et inversement et enfin les comportements dans leur ensemble). Pour permettre la modélisation du contexte ils ont créé une base de données en extrayant du corpus toutes les trames de 3 secondes dans lesquelles il y avait une interaction enfants-parents simultanément ou à la suite de. En plus de l'annotation pré-existante ils ont rajouté des méta-comportements pour grouper les données extraites entre-elles : s'il s'agit de vocalisation, sollicitation vocale, de toucher, de sollicitation gestuelle ou de régulations pour les parents, et de vocalisations, de comportements inter-subjectifs, orientation vers les parents, ou de comportement avec un objet en particulier pour les enfants. Utilisant des HMM, ils ont prédit la probabilité d'apparition de chaque pattern et ont calculé leurs statistiques en utilisant des GLMM (g*eneralized linear mixed model*). Une fois ces statistiques obtenues sur les dynamiques des patterns, ils ont pu concevoir le modèle de reconnaissance automatique en utilisant la classification par clustering. Cette approche développée pour cette étude est peu fréquente dans le champ de la recherche.

Nonobstant, c'est ce que propose Théradia en annotant selon un guide d'annotation qui lui est propre (inspiré par la convention ESLO) les scénarios d'interactions eux même composés de plusieurs catégories : les expressions auto-adressées (neutres ou attitudinales, ou avec la volonté d'abandons) et les expressions adressées (rétro-actives : qui montrent la compréhension ou l'incompréhension d'un patient ; les expressions propositionnelles : les demandes d'approbation par exemple). Chaque expression peut-être verbale ou non-verbale, et chaque segment a été annoté de cette manière selon le guide. L'intérêt de ces scénarios d'interaction est de contextualiser la communication et l'interaction afin d'améliorer l'autonomie et le système de reconnaissance d'affects et d'émotions de l'agent conversationnel.

CREATION ET TRAITEMENT DU CORPUS THERADIA-WoZ

1. RÉCUPÉRATION ET ANNOTATION DES DONNÉES

Afin de mettre en place un système permettant de prédire le contexte d'une certaine interaction, il est nécessaire de récolter des données et de les traiter afin de constituer un corpus d'entraînement pour le système de prédiction. Dans le cadre de THERADIA, les données avaient déjà été récoltées en grande partie afin d'alimenter un système de prédiction d'émotions.

1.1. Protocole THERADIA

Les données du corpus THERADIA-WOZ ont été obtenues via un Magicien d'Oz.

Des sujets, recrutés sur base de volontariat, ont accepté de participer à une ou deux séances avec un assistant virtuel nommé Suzie. L'assistant virtuel était contrôlé par un être humain à distance, conformément à un protocole Wizard of Oz (ou Magicien d'Oz). De fait, les sujets pensaient être face à une véritable intelligence artificielle autonome. Les sujets ont été enregistrés et filmés tout au long de la silence.

En général, une séance dure environ une heure. Durant cette séance, les sujets interagissent avec l'assistant virtuel (ils font connaissance en début de séance, discutent, réagissent à leurs performances, sont rassurés par Suzie, etc...) ou bien exécutent des exercices de mémoire et de logique.

	Jeunes	Séniors	MCI	Total
Nbr de sujets	52	52	9	113
Nbr de sessions	82	71	9	162
Durée total (h)	88h55m11s	79h54m24s	5h59m09s	133h09m56s
Durée moyenne (s)	59m13s	1h07m33s	39m54s	1h00m32s
Durée écart-type (s)	10m28s	11m35s	11m45s	12m35s

Tableau 3: Taille du corpus THERADIA-Woz, en vert les données utilisées

Les sujets sont répartis en trois catégories : une première partition du corpus est celle des seniors, une seconde celle des jeunes adultes, et une dernière, beaucoup plus petite, celle des personnes atteintes de maladie neurodégénérative comme Alzheimer (nommée MCI). Au

total, il s'agit de 78 vidéos qui ont été retenues, annotées et traitées. Les vidéos Jeunes n'ont pas encore été traitées et deux vidéos appartenant à la population des Seniors étaient incomplètes à cause d'un problème d'enregistrement.

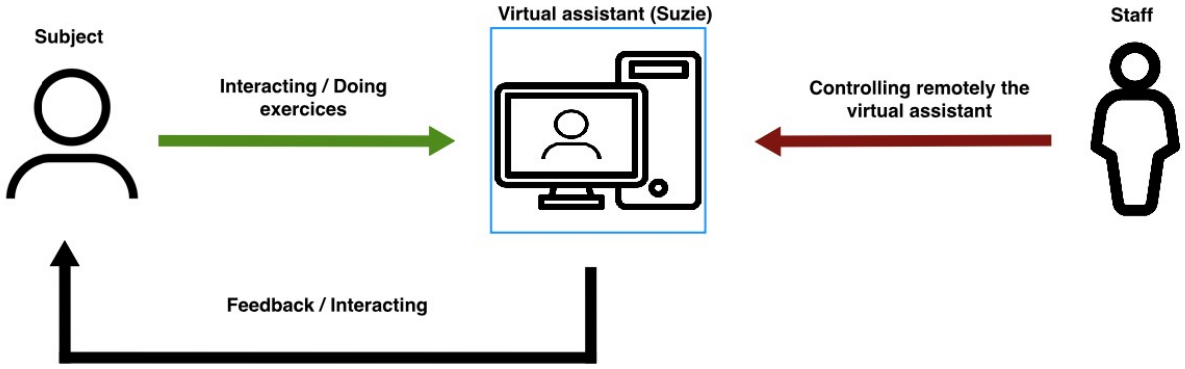

Figure 1: Schéma du protocole Théradia Wizard of Oz (dit Magicien d'oz)

1.2. Segmentation et transcription

Avant que la récolte des données ne soit terminée, le travail de segmentation avait déjà commencé pour gagner du temps. Les vidéos ont été segmentées, transcrites et annotées par plusieurs étudiants. Le travail de segmentation, transcription et d'annotation s'est effectué avec le logiciel libre de droit ELAN[34].

Toutes les vidéos ont été segmentées selon les propositions exprimées par le sujet. Dès lors qu'ils s'expriment de manière verbale ou non-verbale en articulant une même idée, un segment est créé. Ce segment servira ensuite à la transcription.

La transcription suit en grande partie la convention de transcription ESLO.

- Pour l'orthographe et l'existence des mots, la référence était Le Petit Robert. La seule ponctuation conservée était le point d'interrogation.
- Aucune majuscule (hors noms propres) n'a été conservée, toutes les guillemets ont été supprimées, et les apostrophes ne sont conservées qu'en cas d'usage orthographique.

- En ce qui concerne les élisions, seul le *Schwa* est conservé tel quel s'il ne peut-être assimilé à "euh".
- Si une expression est déformée et qu'elle se trouve dans le lexique, alors elle est retranscrite. Sinon, elle est rétablie. Également, les aphérèses sont rétablies afin de faciliter la compréhension.
- Une liste d'onomatopées et d'interjections a été créée et augmentée au fur et à mesure de la transcription.
- Enfin, plusieurs diacritiques ont été créés afin de préciser certaines annotations :

  – <di> Marque le début de l'interaction dans les vidéos.

  – <fi> Marque la fin de l'interaction dans les vidéos.

  – <?> Est placé s'il y a un doute sur la segmentation et la transcription. Après discussion, cette balise est retirée.

  – <nv> Signifie "Non-verbal", est utilisée s'il y a une expression non-verbale dans un segment.

La transcription s'est faîte de manière empirique, évoluant en fonction de la nature des données afin de préciser la convention de transcription. De cette manière, la transcription s'est vue personnalisée afin de correspondre le plus possible à la tâche demandée.

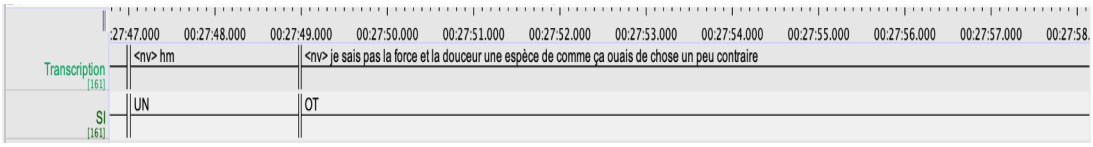

Figure 2: Exemple d'annotation sur ELAN

1.3.Labelisation des données

Lorsque la segmentation et la transcription sont terminées, les segments sont automatiquement extraits avec leurs transcriptions. Chaque segment a été labellisé en fonction de plusieurs critères : nature de l'émotion exprimée ou de l'expression, valence et intensité.

Les segments de chaque vidéo ont été extraits, puis envoyés sur un serveur. Les annotateurs se rendaient sur une page web et regardaient à plusieurs reprises différents segments sélectionnés aléatoirement afin de les évaluer. De cette manière, plusieurs point de vue ont pu être adoptés sur un seul segment. Enfin, il suffisait d'effectuer la moyenne des évaluations sur un segment afin d'obtenir le label final.

1.4. Rajout des Scénarios d'interaction

Ultérieurement, les données ont été augmentées d'une autre piste d'annotation : les scénarios d'interaction. La taxonomie a été créée par Maud Costa (vacataire UGA, Sciences du Langage). Elle est fondée sur les données, en regroupant des segments similaires et en les caractérisant.

Les scénarios d'interaction sont des descripteurs de l'interaction, en d'autres termes ils précisent la nature de l'interaction. Cette interaction est de deux sortes : soit le sujet se parle à lui-même, soit il s'adresse à l'assistant virtuel. Ces deux cas présentent plusieurs spécificités qui ont été décrites afin de préciser l'annotation.

L'objectif des scénarios d'interaction sur le long terme est de proposer une contextualisation des interactions, afin de détecter avec davantage de précision les émotions et attitudes d'un sujet.

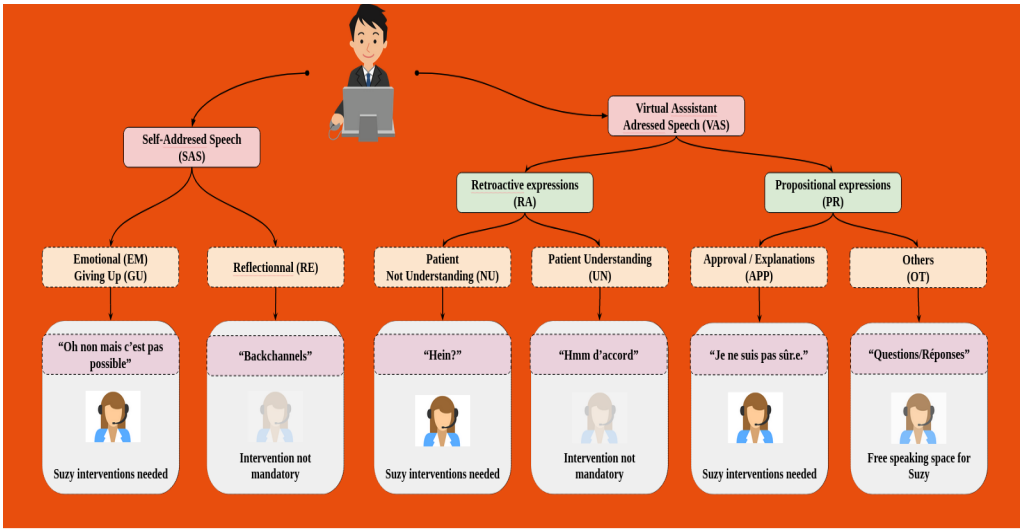

Figure 3: Taxonomie des scénarios d'interaction et interventions de Suzie

1.4.1. Discours auto-adressé - Self-Adressed Speech (SAS)

Dans la grande majorité des cas, le sujet s'adresse à lui-même lors de l'étape de réalisation d'un exercice. Il peut s'agir d'expressions verbales (par exemple le sujet peut s'auto-questionner sur ce qu'il est en train de faire) ou bien de manifestations non-verbales (il exprime une émotion telle que la surprise lors de la découverte d'un exercice ou la difficulté de ce dernier).

Dans le cas où le sujet se parle à lui-même, l'on retrouve trois cas de figure :

- Une attitude réflexive durant laquelle aucune émotion ne transparaît, elle est notée RE pour *Reflexive*.
- Une attitude durant laquelle l'on perçoit une expression émotionnelle positive ou négative, elle est notée EM pour *Emotional*.
- Une attitude qui laisse transparaître la volonté d'abandon du sujet (explicite ou implicite), elle est notée GU pour *Giving Up*.

1.4.2. Discours adressé à l'assistant virtuel - Virtual Assistant Adressed Speech (VAS)

Lors d'interaction avec l'assistant virtuel, deux cas de figures sont possibles selon la taxonomie que nous avons adoptée. Soit le sujet s'adresse à l'assistant virtuel de manière retro-active, soit il s'adresse à elle de manière propositionnelle. Ces deux cas rassemblent l'ensemble des types d'interaction possible en situation de dyade avec l'assistant virtuel.

1.4.2.1. Discours retro-actif - Retro-active Speech (RA)

Une interaction retro-active signifie que le sujet a entendu l'assistant virtuel ou bien lu la consigne d'un exercice. Une fois cette action effectuée, il répond de manière verbale ou non-verbale à l'assistant virtuel. Elle est dite rétro-active car l'interaction constitue une réponse directe à une action passée. Dans ce cas précis, deux annotations ont été prévues :

- Une attitude qui dénote la compréhension, notée UN pour *Understanding*.

- Une attitude qui dénote l'incompréhension, notée NU pour *Not Understanding*.

1.4.2.2. Discours propositionnel - Propositional Speech (PR)

Une interaction propositionnelle est une interaction durant laquelle le sujet questionne ou répond à l'assistant virtuel. Ce type d'interaction est fréquent en début et fin de séance, lorsque l'assistant virtuel questionne le sujet sur sa vie (est-il à la retraite, a-t-il des enfants, a-t-il bien réussi tel ou tel exercice ou encore lequel de ces exercices a-t-il préféré ?).

- Le sujet cherche à être rassuré ou demande l'approbation de l'assistant virtuel, l'attitude est notée APP pour *Approuval*. Il demande indirectement si ses réponses sont bonnes, si l'assistant virtuel peut l'aider, etc.
- Une catégorie "Autre", qui rassemble plusieurs cas de figure, le sujet se fige pour écouter l'assistant virtuel, répond à des questions diverses, discute. Cette catégorie est notée "OT" pour *Other*.

1.5. Uniformisation, vérification et mise à jour des données

Une fois que les données Seniors ont été transcrites et annotées avec les scénarios d'interaction, plusieurs scripts python ont été développés afin de normaliser et de vérifier le contenu des données.

Cette normalisation est devenue une étape obligatoire dès lors que les différentes personnes qui sont intervenues sur les données avaient chacune leurs manières de travailler. Le guide de transcription s'étant adapté et vu augmenté au fur et à mesure des transcriptions, il contenait toujours des zones grises sur lesquelles personne n'avait encore statuées. Le fait de renommer les acteurs, de re-définir les balises de début et de fin de vidéo et de normaliser cela sur l'ensemble des données permet de faciliter l'utilisation de script et l'extraction d'informations dans les étapes ultérieures. Qui plus est, certaines données, ayant été traitées au début de l'étape de transcription en utilisant une première version du guide de transcription

n'étaient plus aux normes quelques mois plus tard lorsqu'une mise à jour du guide de transcription avait été décidée.

La normalisation s'est aussi faite sur les scénarios d'interactions. Une même personne est à l'origine de la première taxonomie des scénarios d'interaction. Néanmoins cette personne a arrêté ce travail avant qu'il ne soit continué par d'autres. Les héritiers de cette taxonomie l'ont parfois mal-comprise et l'étape de vérifications des SI a permis d'uniformiser les données sous une seule et même compréhension de la taxonomie. Également, à l'instar du guide de transcription, la taxonomie des scénarios d'interaction a été amenée à changer. D'une part car la première taxonomie était entièrement rédigée en français et que l'on souhaitait adopter une approche internationale et d'autre part car certaines dénominations se chevauchaient et qu'il était parfois difficile pour un annotateur de faire la part des choses. Ces problèmes ont été réglés dans la mesure du possible tout au long du travail d'annotation et jusqu'à la vérification finale.

Séquences transcrites	Nombre de séquences	Durée totale	Durée moyenne	Durée écart-type
Séniors	14788	36h59m16s	9.00 s	7.82 s
MCI	1105	2h31m44s	8.24 s	5.03 s
Total	15893	39h31m00s	8.95 s	7.66 s

Tableau 4: Nombres de segments et leurs durées

Après la vérification des données, nous obtenons un corpus composé de 15893 transcriptions et annotations, d'une durée totale de 39 heures. Dans le cadre des expériences à venir, nous avons choisi de n'utiliser que les données Seniors, étant un corpus plus vaste et homogène et afin d'éviter d'éventuels biais de résultats.

2. EXTRACTION ET TRAITEMENT DES DONNÉES

2.1. Extraction d'annotations

Une fois que les données ont été vérifiées, les segments sont extraits via un script recueillant le numéro de l'annotation, la transcription corrigée du segment, ainsi que les différents niveaux de scénarios d'interactions du segment en question. Tous les segments ont ainsi été sauvegardés dans un fichier .CSV contenant les différents segments transcrits d'une séance. Chaque segment y est restitué chronologiquement. Par la suite, les timecodes de début et de fin de proposition ont été ajoutés.

À la fin de cette opération est obtenu un fichier .CSV par séance, contenant toutes les propositions numérotées. Dans les étapes suivantes, l'ensemble de ces fichiers sera compilé, afin d'obtenir un corpus contenant dans l'ensemble 15893 propositions. Ce corpus se verra retirer toutes les données concernant la population MCI. Il restera donc 14788 propositions appartenant à la population Seniors.

C'est également grâce à cette étape que nous avons pu obtenir des statistiques préliminaires sur les scénarios d'interactions en calculant leurs fréquences sur l'ensemble du corpus Seniors et MCI.

	Total IS	Total SAS	Total RE	Total EM	Total GU	Total VAS	Total RA	Total NU	Total UN	Total PR	Total APP	Total OT
Total :	15893	7834	5785	2029	20	8059	3355	130	3225	4704	40	4664
Fréquence :		0,492921412	0,36399673	0,12766627	0,00125842	0,507078588	0,0024539	0,0081797	0,202919524	0,29597936	0,002516831	0,29346253

Tableau 5: Nombre total de Scénario d'Interaction (IS) et leurs fréquences sur l'ensemble du corpus

À travers ces statistiques préliminaires, on peut observer une relative représentation uniforme des scénarios d'interactions SAS et VAS (moins de 1 % de différence dans la fréquence d'apparition). Néanmoins, plus l'on précise la nature des interactions plus l'on constate une disparité dans la fréquence d'apparition de certains types de scénarios d'interaction. Nous pouvons remarquer par exemple que parmi les scénarios d'interaction de type SAS, les GU n'apparaissent qu'une vingtaine de fois. Idem pour les scénarios d'interactions NU et APP qui sont sous-représentés au regard de leurs homologues UN et OT. Cette sous-représentation est avant tout liée à la nature du scénario d'interaction, tel qu'elle a été définie. Pour ces trois formes d'interactions, il s'agissait de cas extrêmement rare et dont nous ne pouvions prévoir la fréquence avant de terminer le travail d'annotation des séquences. C'est pour ces raisons que les expérimentations futures ne prendront majoritairement pas en compte ces scénarios d'interaction, pour ne pas fausser les résultats.

2.2. Extraction Audio et Vidéo

Une fois que nous avions les timecodes pour chaque proposition il restait à extraire chaque segment vidéo correspondant à chaque proposition. Pour cela, nous avons utilisé le logiciel FFmpeg[35]. Ainsi, de vidéos d'une heure en moyenne nous avons obtenus plusieurs

clips vidéos dont la durée varie entre deux secondes et plusieurs dizaines de secondes. La tâche étant assez longue, nous avons opté pour une conversion de format, en passant de .MOV à .mp4. L'objectif de cette conversion était également de palier un problème qui survenait lors de la création des segments vidéos. Pour certains clips, nous n'avions que l'audio et non les images dans les premières millisecondes. Cela aurait également faussé les résultats d'openFace quant à la reconnaissance faciale.

Enfin, l'audio a été extrait à partir des segments vidéos obtenus durant l'étape précédente. Tous les clips audios ont été extraits sous le format .wav avec une fréquence d'échantillonnage de 16 kHz pour faciliter l'extraction de caractéristiques acoustiques (MFB, MFCC).

2.3. Traitement openFace

OpenFace est un modèle de reconnaissance faciale disponible gratuitement en ligne[36]. C'est un modèle relativement facile à utiliser permettant d'extraire les unités faciales d'action (Facial Action Units). Le modèle, constitué de réseaux neuronaux profonds, détecte les contours du visage et applique une grille de point sur ce dernier. De cette manière, il est capable d'identifier les différentes parties du visage et de repérer celles qui s'activent image par image. OpenFace repère 17 unité d'actions (sur les 46 composant le FACS – *Facial Action Coding System*) situées au niveau des sourcils, des yeux, du nez et des lèvres puis produit en sortie deux informations : l'intensité de l'activation d'une unité d'action (comprise entre 0 et 5) ainsi que l'activation ou non de l'unité d'action (booléen). Également, openFace produit un score de confiance sur sa capacité à détecter les visages (compris entre 0 et 0.98). Ce score de confiance est utile puisqu'il permet de s'assurer de la validité des données produites. Toutes ces informations sont enregistrées dans un fichier .CSV.

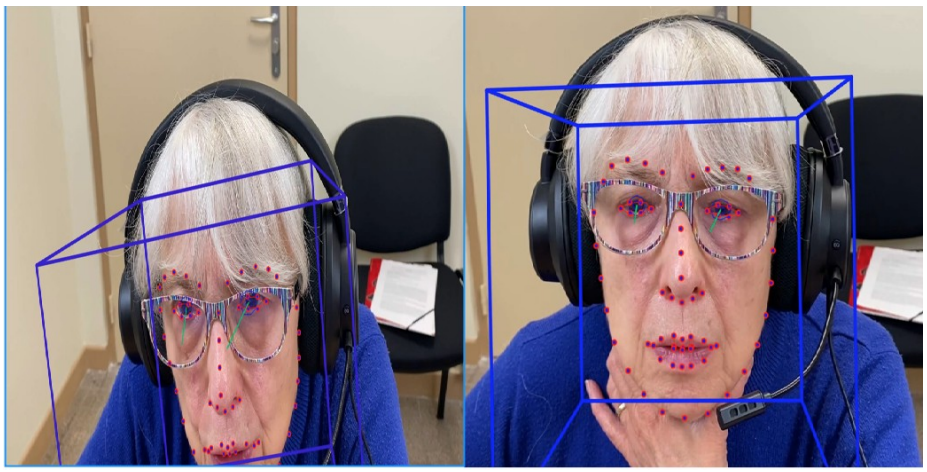

Figure 4: Capture d'écran du tracking d'openFace en fonction du score de confiance. À gauche, la précision estimée d'openFace est de 47%, le sujet est proche de l'écran, le modèle ne peut tracker tout son visage. À droite, la précision d'openFace est estimée à 97%. Le visage est capturée dans son ensemble.

Le tableau 2 liste l'ensemble des Unités d'Action Faciales extraites par openFace et les explicite. On peut constater qu'openFace cible particulièrement les unités d'actions se situant aux alentours des yeux et de la bouche. Les descriptions des Unités d'Action peuvent être retrouvées sur le site *Imotion* (Farnsworth , 2022)[37].

Tableau 6: Liste des FAU's extraites par openFace

AU01	Élévation des sourcils intérieurs
AU02	Élévation des sourcils extérieurs
AU04	Abaissement des sourcils
AU05	Haussement des paupières
AU06	Haussement des joues
AU07	Plissement des paupières
AU09	Nez froncé
AU10	Haussement lèvre supérieure
AU12	Haussement commissure des lèvres
AU14	Activation des fossettes
AU15	Abaissement commissures des lèvres
AU17	Haussement du menton
AU20	Étirement des lèvres
AU23	Resserrement des lèvres
AU25	Légère ouverture des lèvres
AU26	Mâchoires tombantes
AU45	Clignement des yeux

2.4. Création des jeux de données

Les données obtenues en fin de traitement doivent rassembler les numéros de propositions, les transcriptions ainsi que les différents scénarios d'interaction et caractéristiques. openFace créant des .CSV images par images, plusieurs approches ont été adoptées afin de synthétiser le nombre de caractéristiques produites.

- Une première approche consiste à calculer pour chaque unité d'action la moyenne d'intensité sur l'ensemble du segment vidéo. Pour une proposition, on obtient une valeur unique pour chaque unité d'action. De la même manière, on calcule la moyenne du score de prédiction sur l'ensemble de la séance afin d'obtenir une seule valeur.
- La seconde approche consiste à ne récupérer que les valeurs maximales pour chaque unité d'action. Indépendamment, on calcule la moyenne du score de prédiction sur l'ensemble du segment vidéo. Comme pour la première approche, le but ici est de n'obtenir qu'une valeur par unité d'action.
- Ultérieurement, l'amplitude a également été calculée afin de varier les caractéristiques lors de la modélisation des SI via les SVM.

Ces deux approches sont à l'origine de différents types d'ensemble de données qui seront utilisés lors des expériences, un contenant des valeurs moyennes, un autre contenant des valeurs maximales, enfin un troisième contenant l'amplitude. À ces valeurs, l'on ajoute pour chaque proposition le numéro de la proposition, les différents scénarios d'interaction ainsi que la transcription. Enfin on obtient un fichier .CSV contenant toutes les propositions d'une séance ainsi que leurs caractéristiques. Si un sujet a effectué deux séances, alors les fichiers correspondant aux deux séances seront fusionnés afin d'obtenir un seul .CSV par locuteur.

3. LIMITES DES DONNÉES

La première mise en garde à faire sur le système est la répartition des données. Pour le moment, les données jeunes n'ont pas encore été entièrement traitées. Le système s'entraînera donc sur les données appartenant aux partitions des Seniors. Si THERADIA a pour but de

créer et d'entraîner une intelligence artificielle pour un public majoritairement âgé, il est bon de garder en tête que les résultats obtenus dépendent des données. Aussi, il faut attendre que les données de la partition jeune soient traitées et d'observer la différence de résultat une fois qu'elles sont utilisées pour l'entraînement du système.

On peut également s'attarder sur la taxonomie des SI. Plusieurs problèmes sont inhérents au fait d'annoter des types d'interactions. Un premier est lié au format d'enregistrement. Lors des séances, seul le sujet était enregistré. Cela signifie que l'on ne garde que la vidéo et l'audio du sujet durant la séance sans savoir quand l'assistant virtuel s'exprime ou quand un exercice est en cours. Dans le cadre de THERADIA, un autre type de vidéo a été enregistré où l'on voit à la fois le sujet et l'assistant virtuel. Néanmoins, seule la vidéo sans le son a été enregistrée, et la durée des vidéos et différentes de celles où seul le sujet est enregistrée. Leurs utilisations pour étiqueter les SI est donc limitées et ne permet que de clarifier certains doutes lors de l'annotation.

Un autre soucis tient de la complexité des SI. Si la convention est simplifiée, cela permet d'étiqueter plus facilement et d'offrir une vision générale des types d'interactions. Dans les faits, encore beaucoup de types d'interaction se chevauchent dans le contexte d'une interaction réelle. À titre d'exemple, il arrive parfois que les sujets s'adressent à eux-même lorsqu'ils discutent avec l'assistant virtuelle. Dans ce cas de figure, il aurait été juste d'étiqueter l'interaction comme étant une expression auto-adressée, or elle est étiquetée en tant qu'expression adressée à l'assistant virtuel. Néanmoins, il est particulièrement difficile de trancher définitivement et de savoir si cette expression est destinée à être entendue malgré tout ou non. La taxonomie des SI utilisée ici à titre expérimentale simplifie ces questions d'annotations, quitte à perdre parfois en information (contrairement à une taxonomie plus complexe et explicitant les cas de figure possibles). Il est bon de rappeler que le guide d'annotation des scénarios d'interaction a également évolué de manière empirique en fonction des données collectées.

Un dernier reproche que l'on pourrait faire au protocole d'annotation des scénarios d'interactions découle également de la méthodologie utilisée. Afin de simplifier l'extraction d'annotations et d'informations, il a été convenu qu'un segment transcrit serait un seul scénario d'interaction. Un segment peut durer plusieurs secondes (voire dizaines de secondes) et il apparaît naturel que plusieurs SI puissent être présent dans un segment (il peut contenir à

la fois une attitude réflexive et une attitude chargée en émotion). Très tôt deux manières de répondre à cette question ont été adoptées. Soit un scénario d'interaction était majoritairement présent sur un segment, dans ce cas précis, il est utilisé pour décrire le segment, soit un scénario d'interaction est particulièrement intense dans le segment, en dépit du reste du segment qui peut contenir d'autres scénarios d'interaction, dans ce cas, il est sélectionné (par exemple, lors d'un segment d'une vingtaine de secondes, si le sujet adopte une attitude réflexive tout au long du segment mais qu'il l'interrompt en manifestant très fortement une émotion telle que la colère ou la frustration, alors on décide de privilégier l'intensité à la fréquence).

EXPÉRIMENTATIONS SUR LE CORPUS THERADIA-WoZ

1. STATISTIQUES PRÉLIMINAIRES

Avant de commencer les expérimentations et d'entraîner des modèles prédictifs, des mesures statistiques ont été effectuées sur les jeux de données que nous avions obtenus. Ces mesures permettent de former des premières remarques sur les différentes populations que nous souhaitons observer ainsi que des hypothèses sur les résultats des modèles prédictifs (cf. 2. Modélisation des scénarios d'interaction).

1.1. Protocole

Toutes les analyses statistiques ont été effectuée sur la population Senior afin d'éviter d'éventuels biais de résultats (l'on émet l'hypothèse que les sujets MCI adopteront des comportements différents des Seniors or leurs données sont en faible nombre).

Afin d'obtenir des informations pertinentes sur le rôle des Unités d'Actions, nous avons trié les différents clips vidéos par label. Une fois cette première tâche effectuée, la moyenne et l'écart-type de l'intensité de chaque Unité d'Action ont été calculés par locuteur, puis nous en avons fait une moyenne globale en recalculant l'écart-type. Cette première approche nous a permis d'observer les différences ainsi que la dispersion de l'intensité par label.

Une seconde approche a été de prélever la valeur maximale de l'intensité des Unité d'Action pour chaque segment vidéo puis de calculer la moyenne par locuteur, avant de la recalculer sur l'ensemble du label et d'y mesurer la dispersion des valeurs. Cette seconde approche permettra d'observer les différences en mesurant d'abord les valeurs les plus hautes pour chaque Unité d'Action au sein d'un label.

Chacune de ces approches a été répétée trois fois sur trois groupes de données : une première fois sur l'ensemble des données (soit les données Adultes et MCI), une seconde fois uniquement sur les données Seniors et une dernière fois sur les données MCI.

Au vue des résultats obtenus (relativement homogènes), nous avons passer les différentes valeurs obtenues au crible d'un test de Student ($p < 0,05$) pour chaque Unité

d'Action. Ce test se révèle utile afin de comparer la moyenne des populations et d'affirmer qu'une différence entre deux moyennes est significative. Afin de parachever les vérifications, nous avons également effectué des tests d'Anova (*one-way anova*, $p < 0.05$) sur l'ensemble des valeurs afin d'observer la variance entre les valeurs obtenues par Unité d'Action pour chaque label. Ces deux tests nous permettent d'établir que pour chaque Unité d'Action, les moyennes obtenues par label seront bien différentes et significatives.

1.2. Statistiques sur les scénarios d'interaction du premier niveau

La première batterie de tests statistiques a été effectuée sur les scénarios d'interaction (SAS / VAS) afin de savoir si le sujet s'adresse à lui-même ou bien à l'assistant virtuel Suzie.

1.2.1. Comparaison des valeurs d'intensité à partir des moyennes

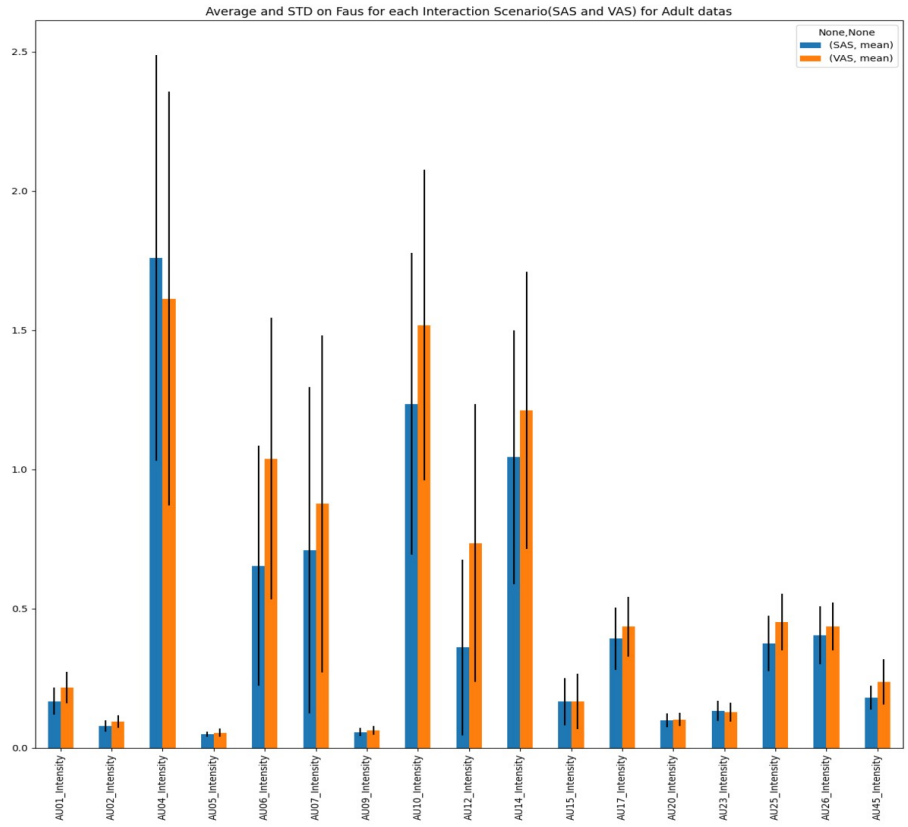

Figure 5: Valeurs d'intensité des UA par label (SAS / VAS) sur les données Senior

Par l'approche statistique fondée sur les moyennes des valeurs d'intensité, on constate qu'une poignée d'Unité d'Action (UA) se démarque par leurs intensités et donc leur activation. Dans l'ordre, il s'agit des UA04 (abaissement des sourcils) , UA06 (haussement des joues), UA07 (plissement des paupières), UA10 (haussement lèvre supérieure), UA14 (activation des fossettes). Parmi elles, on observe que les UA06, UA07, UA10, et UA14 sont plus souvent intenses lorsque le sujet s'adresse à l'assistant virtuel. Également, UA12 semble tenir un rôle important afin de discriminer SAS et VAS. Néanmoins, les écart-types (la dispersion des valeurs autour de la moyenne) est similaire pour chaque UA. À fin de discerner si les populations sont réellement différentes un t-test a été effectué.

AU	Résultats Anova	P-Value	Résultats T-test	P-Value
AU01_Intensity	35.156979	1.928438e-08	-5.316838827277688	4.220667536055137e-07
AU02_Intensity	23.023666	3.750717e-06	-4.280096282127167	3.502239992053779e-05
AU04_Intensity	1.604294	2.072082e-01	1.1796811522403747	0.24018650030352642
AU05_Intensity	4.819198	2.964109e-02	-2.090584772312968	0.03842650690682498
AU06_Intensity	23.530280	2.983864e-06	-4.802819249663001	4.073180231887149e-06
AU07_Intensity	2.515398	1.147904e-01	-1.617885853392681	0.10800418903880377
AU09_Intensity	6.375250	1.258422e-02	-2.2221377491661674	0.027926168903377895
AU10_Intensity	9.815891	2.071515e-03	-2.9948820625141774	0.0032630121637707006
AU12_Intensity	32.477263	5.964407e-08	-5.275004180270287	5.102540385498895e-07
AU14_Intensity	4.871213	2.878788e-02	-2.0601288965141107	0.04129282515685722
AU15_Intensity	0.006461	9.360415e-01	-0.0835741113825057	0.9335179403657398
AU17_Intensity	5.729940	1.788137e-02	-2.2789756124794778	0.024226184719617277
AU20_Intensity	0.882312	3.490401e-01	-0.7545364401296195	0.4518317304471551
AU23_Intensity	0.241216	6.240281e-01	0.7393533237483623	0.46096670703794074
AU25_Intensity	25.695339	1.133020e-06	-4.507930226143688	1.398924373207851e-05
AU26_Intensity	4.985118	2.700845e-02	-1.9374122792559316	0.054767700381264856
AU45_Intensity	25.611198	1.176152e-06	-5.088337610839092	1.1767024141386296e-06

Tableau 7: Résultats des tests Anova et T-test ($p < 0.05$) sur les valeurs moyennes d'intensité des UA pour les labels SAS et VAS. En rouge, le test n'est pas significatif.

Grâce au t-test, on peut observer que les UA04 et 07 qui étaient particulièrement discriminantes en ne tenant en compte que la moyenne se révèlent être non-significatives on ne peut donc conclure que ces UA soient d'une importance capitale dans la détection de contexte. En revanche, l'UA14 dont l'intensité élevée permettait d'en faire une Unité d'Action référente des VAS se trouve validée par le test statistique.

Pour les SI de premier niveau (SAS / VAS), le test de l'anova semble confirmer les résultats du T-test. Les UA04, 07, 15, 20 et 23 ne sont pas significatives tout comme pour le T-test. En revanche, ce test semble affirmer que l'UA26 est un critère discriminant. Cela nous contraint à observer que l'on ne peut statuer clairement sur le rôle que joue cette Unité d'Action lors de la détection de contexte.

1.2.1 Comparaison des valeurs d'intensité à partir des valeurs maximales

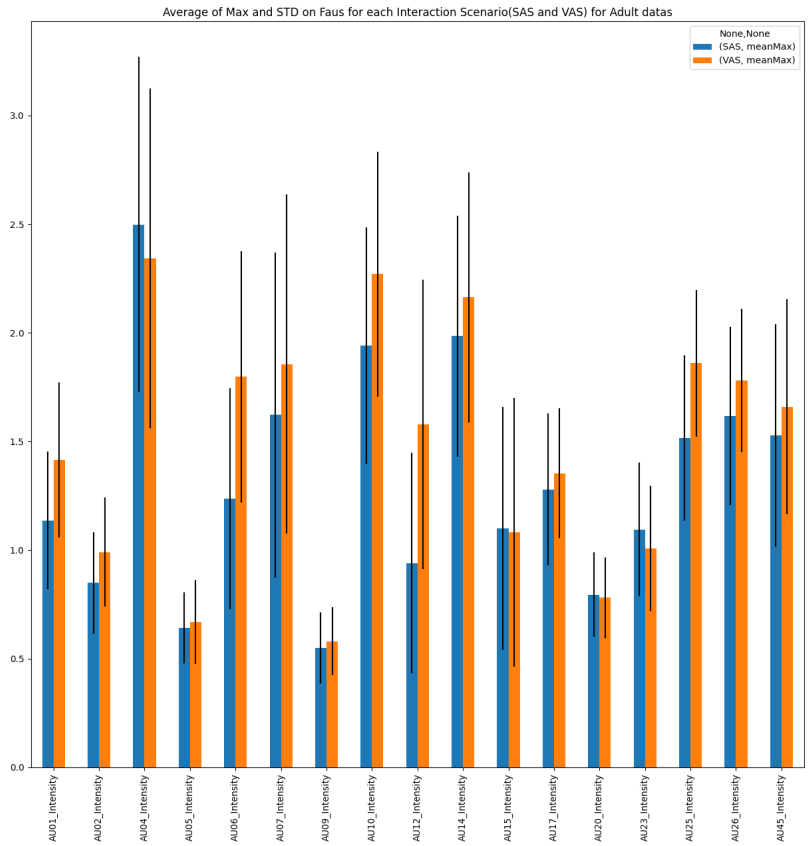

Figure 6: Valeurs d'intensité des UA par label (SAS / VAS) sur les données Adultes à partir des valeurs maximales

On observe avec cette seconde approche une distribution similaire à celle des moyennes. Les UA04, UA06, UA07, UA10, UA14 ont tendance à être les plus intenses. Néanmoins les valeurs maximales tendent à montrer l'importance des UA25 (légère ouverture des lèvres) et 26 (mâchoires tombantes) lorsque le sujet s'adresse à l'assistant virtuel. On constate que l'écart-type de ces UA est légèrement inférieur lorsque l'interaction est labellisée

VAS tout en ayant une plus forte intensité lors de l'activation. De la même manière, UA12 semble caractéristique du scénario VAS.

AU	Résultats Anova	P-Value	Résultats T-test	P-Value
AU01_Intensity	23.420554	3.481723e-06	-4.840336722009975	3.4689295900044032e-06
AU02_Intensity	11.835805	7.719535e-04	-3.4327272338570394	0.0007921068491821545
AU04_Intensity	1.396344	2.393988e-01	1.1870676878652362	0.23727024343269976
AU05_Intensity	0.795562	3.739991e-01	-0.8771551609786029	0.38194920445693625
AU06_Intensity	36.541054	1.356331e-08	-6.040755407696236	1.3842121433911804e-08
AU07_Intensity	3.238815	7.412956e-02	-1.77455077248396	0.07820943445318411
AU09_Intensity	1.303898	2.555090e-01	-1.1499444384737905	0.25218486415271113
AU10_Intensity	12.153115	6.602206e-04	-3.462709265230414	0.0007152974440665679
AU12_Intensity	40.057564	3.325678e-09	-6.327960874986441	3.3447644941964244e-09
AU14_Intensity	3.478732	6.431719e-02	-1.8610148827276642	0.06490117831078195
AU15_Intensity	0.031675	8.590077e-01	0.1793963942210902	0.85789346228739
AU17_Intensity	1.765039	1.862222e-01	-1.3269929727910603	0.186734612665152
AU20_Intensity	0.170592	6.802367e-01	0.42142337146207637	0.6741112690130568
AU23_Intensity	2.896762	9.104250e-02	1.74732424503194	0.08283831373008001
AU25_Intensity	31.628695	1.020560e-07	-5.600969074611283	1.136697115067129e-07
AU26_Intensity	6.604439	1.124924e-02	-2.5413576234205837	0.012164267602538375
AU45_Intensity	2.368357	1.261408e-01	-1.5362384211554851	0.12680316123058902

Tableau 8: Résultats des tests Anova et T-test ($p < 0.05$) sur les valeurs maximales d'intensité des UA pour les labels SAS et VAS. En rouge, le test n'est pas significatif.

Le T-test effectué sur les valeurs maximales est bien plus souvent non-significatif. En d'autres mots, pour les UA04, UA05, UA07, UA09, UA14, UA15, UA17, UA20, UA23 et UA45, nous n'avons pas pu prouver que les populations étaient radicalement différentes et donc que l'UA pouvait déterminer le contexte. On remarque cependant que les UA06 et UA10 semblent être caractéristiques du VAS (elles sont validées à la fois par le T-test effectué sur valeurs moyennes et le T-test effectué sur valeurs maximales). À l'inverse, l'UA26 récemment discriminante semble être validée par ce T-test, contrairement à celui effectué sur les valeurs moyennes.

On retrouve également dans ce T-test l'ensemble des UA non-significatives du premier t-test, c'est à dire les UA23, 20, 15, 07 et 04. Exceptée l'UA26 dont le statut est ambiguë, il semblerait que ces UA ne permettent pas, à l'heure actuelle, d'infirmer ou affirmer la présence d'un certain contexte avec certitude. Néanmoins, on ne peut tout à fait rejeter l'hypothèse qu'elles peuvent avoir un rôle décisionnaire couplées avec d'autres Unités d'Action.

Tout comme pour pour le t-test exécuté sur les valeurs maximales, davantage d'Unité d'Action ne sont pas significatives au sens où l'on pourrait définitivement affirmer qu'elles joueraient un rôle dans la détection de contexte. L'anova conclue des résultats similaires au t-test en ce qui concerne les UA04, 05, 07, 09, 14, 15, 17, 20, 23 et 45. Les deux tests semblent établir ensemble que ces Unités d'Action jouent un rôle mineur dans la détection de contexte puisque l'on ne saurait à dire si les populations étudiées sont différentes.

1.3. Statistiques sur les scénarios d'interaction de deuxième niveau

Les tests effectués sur le premier niveau d'interaction ont également été réalisés sur les deux groupes de scénarios d'interaction SAS et VAS. Le but visé étant d'observer une fois de plus si à l'intérieur de ces groupes il y aurait des Unités d'Action qui permettraient de reconnaître plus efficacement le contexte.

1.3.1. Statistiques sur les scénarios d'interaction SAS

Le premier groupe SAS se compose de trois scénarios d'interaction : RE si le sujet adopte un comportement réflexif (il relit la consigne, réfléchit à voix basse), EM si le comportement du sujet témoigne d'affect (soufflement, injures etc.) et enfin GU si le sujet évoque l'envie d'abandonner. Les scénario d'interaction EM et GU ont été fusionnés à ce stade de l'étude pour deux raisons : premièrement ils étaient relativement proches dans leurs manifestations, les deux véhiculant des signaux sociaux chargés d'affects, deuxièmement car le scénario d'interaction GU apparaît peu de fois (cf. Tableau 5).

1.3.1.1. Comparaison des valeurs d'intensité à partir des valeurs moyennes

La distribution obtenue par les valeurs moyennes pour chaque Unité d'Action est similaire à celle des scénarios d'interaction de premier niveau (SAS / VAS). C'est un constat logique puisque les valeurs des SAS sont les mêmes que les valeurs des RE et des EM. Également, les valeurs moyennes entre les deux labels sont particulièrement égales. Les UA 04, 06, 07, 10 et 14, restent les plus fréquentes ; néanmoins, ce sont dans certaines Unités

d'Action, moins intenses et fréquentes que l'on retrouve des différences flagrantes.

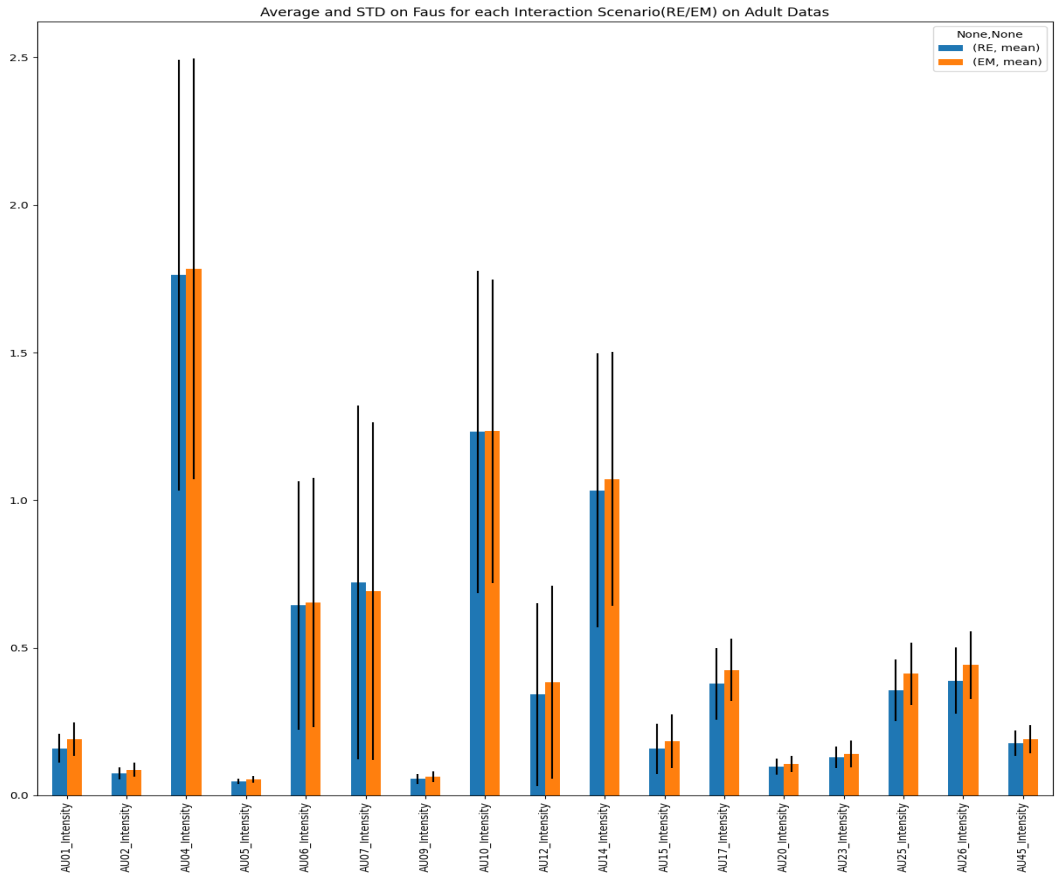

Figure 7: Valeurs moyennes d'intensité des UA par label (RE / EM) sur les données Adultes à partir des valeurs moyennes

Les UA25 et 26 sont légèrement plus actives lors du contexte EM que RE. On ne peut encore considérer ces différences comme étant significatives car l'observation de l'écart-type pour chaque UA nous force à penser que les Unités d'Actions sont peu utiles pour discriminer les scénarios d'interaction RE et EM. Dernier constat, le label RE souffre d'un écart-type généralement supérieur sur chaque Unité d'action malgré les moyennes légèrement inférieures. On peut en déduire que ce contexte sera plus difficile à repérer car il apparaît sous des formes différentes. Les Unités d'actions seules ne permettront pas de le reconnaître efficacement.

AU	Résultats Anova	P-Value	Résultats T-test	P-Value
AU01_Intensity	11.609385	0.000863	-3.439142437503796	0.0007750484293915799
AU02_Intensity	10.023071	0.001908	-3.176434546543014	0.0018451919294570974
AU04_Intensity	0.031596	0.859181	-0.174912765086147	0.8614083127262909
AU05_Intensity	11.057039	0.001136	-3.326245719807219	0.0011320371954371318
AU06_Intensity	0.020011	0.887717	-0.13192756148406454	0.8952365652105363
AU07_Intensity	0.082287	0.774659	0.29507982315186965	0.7683828909065712
AU09_Intensity	7.285944	0.007832	-2.717027451765085	0.007443992249658427
AU10_Intensity	0.003036	0.956142	-0.027734400992682998	0.977914642130146
AU12_Intensity	0.584597	0.445841	-0.7596636558950157	0.4487704543445118
AU14_Intensity	0.259803	0.611081	-0.5043208108202973	0.6148520396123842
AU15_Intensity	3.086073	0.081216	-1.7598291433152913	0.08068516008834661
AU17_Intensity	5.868759	0.016728	-2.41226449849135	0.017187408860865394
AU20_Intensity	3.803771	0.053194	-1.936004920915464	0.05494155388707253
AU23_Intensity	2.773037	0.098166	-1.624131497688921	0.10666322927289452
AU25_Intensity	10.070539	0.001863	-3.1772540426487925	0.001840348823068002
AU26_Intensity	7.554128	0.006800	-2.71665053808291	0.0074520289819852885
AU45_Intensity	3.449538	0.065432	-1.852148968168817	0.06617256445253808

Tableau 9: Résultats des tests Anova et T-test ($p < 0.05$) sur les valeurs moyennes d'intensité des UA pour les labels RE et EM. En rouge, le test n'est pas significatif.

Les T-tests et l'Anova ont permis de souligner que plus de la moitié des Unités d'Action ne sont pas significatives. On ne peut donc affirmer avec certitude que les UA04, 06, 07, 10 et 15, qui étaient particulièrement intenses, permettent de discriminer les deux contextes. En revanche, les UA25 et 26 qui étaient plus actives lors d'un contexte EM apparaissent comme significatives. On peut en déduire qu'elles joueront un rôle lors de la détection de contexte par notre modèle prédictif. Néanmoins, le fait qu'elles soient peu intenses sur l'ensemble du corpus les placent immédiatement comme un facteur de décision secondaire.

1.3.1.2. Comparaison des valeurs d'intensité à partir des valeurs maximales

Les valeurs maximales permettent de mettre davantage en évidence les différences entre les labels RE et EM. L'écart-type des moyennes maximales apparaît plus faible si ce n'est égal lors du contexte RE, mais l'intensité des Unités d'Action est en moyenne plus élevée lors d'un contexte EM. Deux raisons peuvent expliquer cela : d'une part, comme énoncé plus tôt, le contexte RE est particulièrement varié, ce qui explique qu'il ait de manière générale une intensité plus faible, d'autre part le contexte EM est un contexte pendant lequel

le sujet émet des signaux sociaux fortement chargés en affect, il apparaît logique dans ce contexte que l'intensité des Unité d'Action soit plus forte.

AU	Résultats Anova	P-Value	Résultats T-test	P-Value
AU01_Intensity	15.629693	0.000123	-3.9835643774192726	0.0001101256748723610
AU02_Intensity	13.024466	0.000431	-3.6147428804639827	0.0004223194903277854
AU04_Intensity	0.636884	0.426233	-0.7989007543668918	0.4257409454220509
AU05_Intensity	7.780430	0.006040	-2.783236656043877	0.006148323806384846
AU06_Intensity	2.577766	0.110694	-1.5987065942909342	0.11220652458735689
AU07_Intensity	0.366127	0.546131	-0.600034236825704	0.5494818991506794
AU09_Intensity	5.703982	0.018300	-2.3974349595506292	0.01786878056447057
AU10_Intensity	1.494399	0.223650	-1.1980205393917376	0.23299268441215698
AU12_Intensity	4.813744	0.029932	-2.184459938882957	0.030642214837583448
AU14_Intensity	5.157451	0.024718	-2.267521710476069	0.02493520337164852
AU15_Intensity	3.336159	0.069966	-1.8281339234357668	0.06972091793274314
AU17_Intensity	6.210604	0.013900	-2.4871814331149014	0.014085540524611244
AU20_Intensity	3.574939	0.060785	-1.8468148311734083	0.06694743652471503
AU23_Intensity	1.953706	0.164464	-1.3818518916406874	0.16928281196703096
AU25_Intensity	13.131865	0.000409	-3.6233187566172487	0.0004097547783341321
AU26_Intensity	8.268955	0.004683	-2.8649825700854654	0.004833083903155738
AU45_Intensity	1.748982	0.188224	-1.3098121247445578	0.1924676150060111

Tableau 10: Résultats des tests Anova et T-test ($p < 0.05$) sur les valeurs maximales d'intensité des UA pour les labels RE et EM. En rouge, le test n'est pas significatif.

Tout comme pour les tests statistiques effectués sur les valeurs moyennes, le T-test et l'Anova ont ici des résultats analogues. Les UA04, 06, 07, 10, 15, 20, 23 et 45 n'obtiennent pas de P-value inférieure à 0,05 et ne sont donc pas considérées comme significatives. On ne peut encore prouver leur importance lors de la décision de contexte. À l'inverse, les UA 25 et 26, qui tenaient un rôle décisionnaire lors des tests statistiques effectués sur les moyennes sont ici aussi significatives. Qui plus est, elles se révèlent parfois importantes lors du contexte EM puisque la moyenne de leurs valeurs maximales obtient un score d'intensité légèrement supérieur à 1,5. Ces analyses nous permettent de statuer sur la possibilité de repérer efficacement le scénario d'interaction EM contrairement au scénario d'interaction RE.

1.3.2. Statistiques sur les scénarios d'interaction VAS

Le label VAS réunit deux types de scénarios d'interaction. Un premier nommé RA (rétro-actif) qui indique si le sujet a compris ou non la situation qui se présente à lui comme un exercice ou les questions posées par Suzie. Un second type de scénarios d'interaction est

quant à lui nommé PR (Propositionnal). Le groupe interactionnel PR forme un ensemble particulier puisqu'à l'intérieur se situe dans un premier temps le scénario d'interaction APP (lorsque le sujet cherche à être rassuré par Suzie). Dans un second temps, l'on retrouve le scénario d'interaction OT. Ce dernier est particulièrement intéressant puisqu'il s'agit d'un espace ouvert de discussion entre le sujet et l'assistant virtuel Suzie, espace de discussion qui est à même de fournir des expressions faciales hétérogènes.

1.3.2.1. Comparaison des valeurs d'intensité à partir des valeurs moyennes

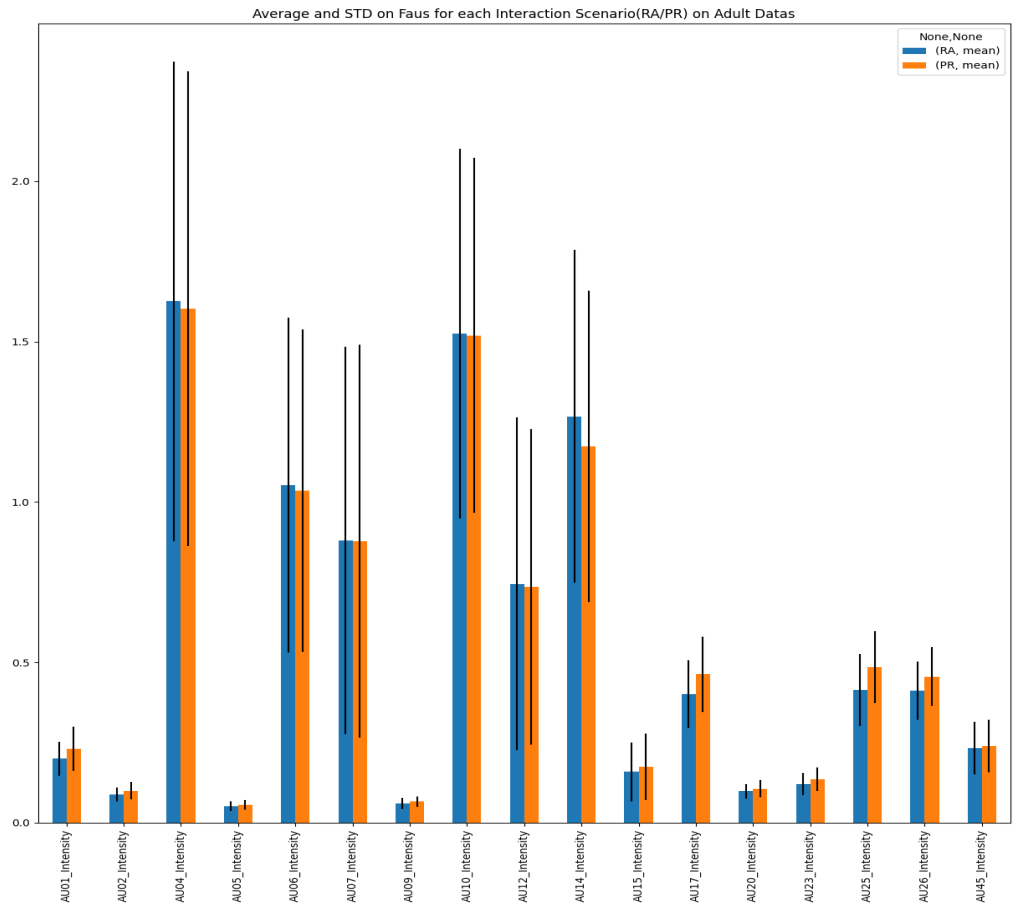

Figure 8: Valeurs moyennes d'intensité des UA par label (RA / PR) sur les données Adultes à partir des valeurs moyennes

À la lumière du graphique, on retrouve les Unités d'Action habituelles (UA04, UA06, UA07, UA10 et UA14) qui composent notre ensemble d'UA les plus fréquentes, les plus

intenses et actives. Les moyennes d'intensité et leurs écart-types sont similaires sur l'ensemble des Unités d'Action, aussi peut-on en conclure que par l'usage des moyennes, il apparaît difficile de reconnaître un contexte ou l'autre. Seules exceptions, les UA17, 25 et 26 dont l'intensité semble bien varier d'un contexte à l'autre. On peut également remarquer que, comparativement aux scénarios d'interaction appartenant au groupe SAS, la moyenne de l'intensité de l'UA 10 est nettement supérieure dans le groupe VAS.

AU	Résultats Anova	P-Value	Résultats T-test	P-Value
AU01_Intensity	8.526389	0.004099	-2.9215952193689505	0.004078936218035546
AU02_Intensity	7.944794	0.005543	-2.8079555293109313	0.005719781222751111
AU04_Intensity	0.031657	0.859047	0.18270447898472	0.8553019768674677
AU05_Intensity	3.350317	0.069381	-1.8390047193622547	0.06809559783553111
AU06_Intensity	0.038845	0.844049	0.19456320295711546	0.846025250695709
AU07_Intensity	0.000649	0.979717	0.028098572193339066	0.9776247230116248
AU09_Intensity	3.970857	0.048296	-1.9868805182292089	0.04894502171942295
AU10_Intensity	0.004162	0.948654	0.0590806826536464	0.9529745685193257
AU12_Intensity	0.015390	0.901453	0.12129878610288446	0.9036334169078919
AU14_Intensity	1.201826	0.274894	1.0914787835766884	0.27699213996718536
AU15_Intensity	0.887211	0.347904	-0.9382419044622679	0.3497833814473563
AU17_Intensity	10.360960	0.001610	-3.212403390717949	0.0016434009448407462
AU20_Intensity	3.654260	0.058030	-1.8903403446100073	0.060840178077653816
AU23_Intensity	5.723923	0.018102	-2.3580299642819265	0.01979691647894662
AU25_Intensity	13.554616	0.000333	-3.6688851246361462	0.0003487055123506075
AU26_Intensity	8.149728	0.004982	-2.836195762662908	0.005263637517262499
AU45_Intensity	0.266265	0.606686	-0.5167704781683577	0.6061557322722675

Tableau 11: Résultats des tests Anova et T-test ($p < 0.05$) sur les valeurs moyennes d'intensité des UA pour les labels RA et PR. En rouge, le test n'est pas significatif.

Les résultats des deux tests convergent sur les différentes Unités d'Actions. Les UA les plus fréquentes et intenses ne sont pas significatives contrairement à celles dont la moyenne d'intensité est inférieure à 1. Les UA 17, 25 et 26 demeurent significatives et permettent d'établir des différences notoires entre le contexte PR et le contexte RA. Tout comme pour le groupe SAS, il est plus difficile d'établir un pattern discriminant entre les deux contextes et l'on peut s'attendre lors de la modélisation à des scores de prédictions relativement faibles.

1.3.2.2. Comparaison des valeurs d'intensité à partir des valeurs maximales

La distribution des résultats obtenue à partir de la moyenne des valeurs maximales est similaire à celle des valeurs maximales du groupe SAS. Les mêmes Unités

d'Action (UA04, 06, 07, 10, 14, 25, 26 et 45) ont un score d'intensité élevé. Si le label PR contient de manière générale des Unités d'Actions qui se démarquent par leurs intensités, on remarque que leurs écarts-types est sensiblement similaires entre les labels RA et PR. Il apparaît donc difficile à la vue du tableau de définir qu'un ensemble d'Unités d'Action caractériserait un label en particulier, donc un contexte.

À l'instar de l'ensemble des tests statistiques effectués sur les scénarios d'interaction de niveau 2, les Unités d'Action les plus fréquentes et intenses ne sont pas significatives. Aussi notre groupe d'UA (UA04, UA07, UA10, UA14, UA45) ne nous permet pas d'établir avec certitude une discrimination entre nos deux labels. En revanche les UA02, 17, 25 et 26 semblent adopter des comportements différents entre les deux labels.

AU	Résultats	P-Value	Résultats T-test	P-Value
AU01_Intensity	12.162901	0.000657	-3.4821940666283444	0.0006691921472305494
AU02_Intensity	12.155852	0.000659	-3.4796703270590346	0.0006750023161876616
AU04_Intensity	0.149349	0.699762	-0.38285040651136426	0.7024283956666422
AU05_Intensity	6.191168	0.014047	-2.494948668628558	0.013794418108625557
AU06_Intensity	1.290093	0.258028	-1.1393846327331196	0.25654559652740117
AU07_Intensity	0.307628	0.580051	-0.5503539295042872	0.5829794546294862
AU09_Intensity	5.453722	0.020990	-2.3200076745388083	0.021829206358434338
AU10_Intensity	0.936063	0.335010	-0.9723629278476371	0.3325962759682144
AU12_Intensity	1.703631	0.194018	-1.3028478874409548	0.19482837809332013
AU14_Intensity	0.037536	0.846667	-0.19725535558941382	0.8439222133710111
AU15_Intensity	1.938874	0.166063	-1.3925449928542328	0.16602995966969988
AU17_Intensity	15.703718	0.000119	-3.9528148949638404	0.00012361426739734713
AU20_Intensity	5.989097	0.015670	-2.4254888908789782	0.016599330658700555
AU23_Intensity	5.187906	0.024304	-2.247897277885335	0.02619227547072997
AU25_Intensity	14.666014	0.000195	-3.820664937188197	0.00020166302445326576
AU26_Intensity	10.489937	0.001509	-3.2315837595284873	0.0015443822545336604
AU45_Intensity	0.873280	0.351705	-0.9366846428254906	0.3505811583424634

Tableau 12: Résultats des tests Anova et T-test (p < 0.05) sur les valeurs maximales d'intensité des UA pour les labels RA et PR. En rouge, le test n'est pas significatif.

1.4. Conclusions statistiques

Ces résultats sont sujets à caution. En premier lieu car l'objectif de ces statistiques étaient de vérifier l'hypothèse *a priori* que les Unités d'Actions se manifestaient de différentes manières en fonction des contextes. Également, les scores d'intensités en fonction des contextes sont extrêmement proches, ce qui laisse relativiser sur l'importance de telle

Unité d'Action au sein d'un contexte donné. Néanmoins, ces analyses ont permis de souligner plusieurs traits caractéristiques des Unités d'Action au sein de notre corpus.

D'abord on peut observer que certaines Unités d'Action ont un score d'intensité particulièrement élevé et ce, fréquemment. Il s'agit des UA 04, 10 et 14 qui correspondent respectivement à l'abaissement des sourcils, au haussement de la lèvre supérieure, et à la contraction des fossettes. Cette régularité en terme d'intensité peut être interprétée de deux manières distinctes : d'une part on ne peut omettre qu'openFace est peut-être sensible à ces Unités d'Action, d'autre part, on pourrait en conclure qu'il y a des Unités d'Action qui s'utilisent dans plusieurs contextes différents. Ce dernier point apparaît comme crucial, puisque les Unités d'Action ont d'abord été pensées pour étudier et interpréter la structure physique des expressions faciales. Or, nous formulons ici l'hypothèse que certaines expressions faciales sont plus fréquentes dans un contexte donnée qu'un autre. openFace nous livre l'intensité des Unités d'Action image par image. Il faudrait annoter les sorties d'openFace pour que nous puissions, à un instant *t* comprendre quelles unités d'action se mobilisent, au sein de quelle séquence et dans quel but.

Ce soucis d'identification d'un contexte par les Unités d'Action rend particulièrement complexe la discrimination des contextes à l'intérieur des labels SAS et VAS. Comme on peut l'observer sur les différentes figures, la distribution des scores d'intensité est relativement similaire entre les labels EM/RE et PR/RA. À ce sujet, on peut observer le nombre de tests considérés comme significatif pour les labels SAS/VAS et les scénarios d'interaction de niveau 2. Cela nous indique que les modèles prédictifs que nous mettrons en place ultérieurement auront très probablement des scores de prédiction bien inférieurs lorsqu'il s'agira d'évaluer leurs capacités à prédire les scénarios d'interaction de niveau 2.

En dernier lieu, considérons les Unités d'Actions significatives par labels chez les scénarios d'interactions de niveau 2. Si l'on compare les résultats des T-test et de l'Anova ainsi que les graphiques d'intensité, on observe qu'à l'intérieur du groupe SAS et du groupe VAS les mêmes Unités d'Action permettent d'établir une différence entre les scénarios d'interaction (les UA 25 et 26 par exemple) . Cela nous amène à penser qu'il est probablement difficile de discriminer les scénarios d'interaction RE, EM, PR, et RA au sein d'une tâche de classification multi-classe.

2. MODÉLISATION DES SI

Une fois le corpus prêt à l'emploi et les analyses statistiques effectuées, les scénarios d'interaction peuvent être modéliser. Dans le cadre du stage, nous avons opté pour des SVM qui devaient au travers des différentes valeurs des FAU être capables de reconnaître un contexte spécifique.

2.1. Protocole

La partie suivante détaillera les différentes étapes ayant précédé la réalisation des modèles prédictifs.

2.1.1. Partitionnement du corpus

Afin d'éviter que les modèles prédictifs soient trop spécialisés sur les données utilisées, nous avons choisi pour approche la validation croisée sur strates. L'objectif derrière étant d'évaluer la capacité du système à prédire correctement sur de nouvelles données mais également d'éviter les biais lors de la prédiction.

Une partition contient cinq itérations de sous-partitions Entraînement, Validation et Test. Chacune des sous-partitions contient respectivement 80 % des données (pour l'entraînement) et 10 % pour la validation et le test. À chaque itération, les données sont mélangées de telle sorte à ce qu'une donnée qui s'est retrouvée dans la sous-partition *Validation* ou *Test* ne s'y retrouve plus.

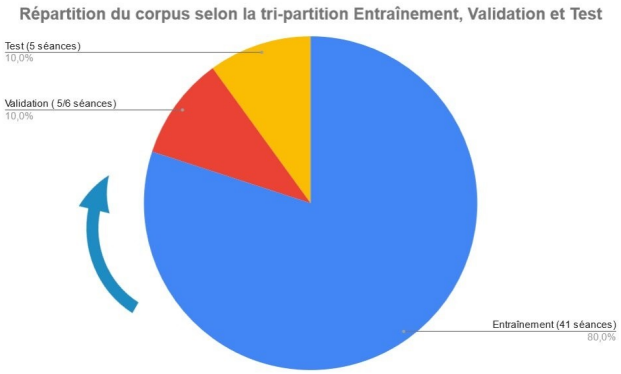

Figure 9: Répartition du Corpus lors de la Validation Croisée à 5 strates

Comme on peut l'observer sur le graphique, les données ont été regroupées par locuteurs (ou sujets) afin d'étudier les caractéristiques propres à chaque sujet. Aussi, la sous-partition d'Entraînement est composée de 41 locuteurs différents, les deux autres partitions en contiennent entre 5 à 6. Si un sujet a déjà fait parti des sous-partitions Validation ou Test il sera affecté à la sous-partition Entraînement pour les prochaines itérations. De cette manière on s'assure que les sous-partitions *Validation* (utilisées pour l'optimisation des hyper-paramètres) et *Test* (utilisées pour l'évaluation du modèle) contiennent toujours de nouvelles données pour le modèle et non des données sur lesquelles il a déjà performé, afin de garantir le panel d'échantillons variés.

À l'instar des analyses statistiques, nous n'utiliserons ici que les données Seniors pour éviter d'éventuels biais de résultats.

2.1.2. Caractéristiques extraites

Afin d'aider les SVM à classifier les données, plusieurs caractéristiques des FAUs ont été extraites dans le but d'alimenter plusieurs modèles.

- *La Moyenne* : la moyenne de chaque FAU a été calculée pour chaque segment vidéo, puis pour chaque locuteur. C'est la première caractéristique extraite car elle permet d'avoir un aperçu général des FAUs pour un label donné.
- *La valeur Maximale* : la valeur maximale de chaque FAU a été extraite pour tous les segments vidéos. Cette caractéristique a été extraite car l'on postulait *a priori* qu'elle permettrait de discriminer plus facilement les labels puisque l'on aurait des valeurs extrêmement disparates entre les FAU pour un label.
- *L'Amplitude* : l'amplitude a été calculée en prenant en compte dans chaque segment vidéo la valeur maximale et la valeur minimale avant de les soustraire. C'est une caractéristique qui dans une moindre mesure permet de rendre compte des FAUs dans leurs distributions.
- *L'Écart-type* : l'écart-type a été calculé sur les trois caractéristiques extraites. Cette caractéristique est importante car elle donne un aperçu de la distribution des valeurs pendant une trame.

- *Score de Confiance d'openFace* : Si ce n'est pas une caractéristique utilisée par les SVM, elle se révèle utile puisqu'elle permet de filtrer le bruit du corpus.

Les SVM sont particulièrement sensibles au bruit, le score de confiance nous a donc permis d'obtenir des résultats sur différents ensembles de données au taux de bruit varié.

Seuil de précision	Entraînement	Validation	Test	Total	Pourcent. / Total
0	11830	1653	1304	14787	100%
85	11486	1586	1286	14358	97,09%
90	11013	1497	1256	13766	93,08%
95	8769	1137	1055	10961	74,12%
96	7831	1012	946	9789	66,19%

Tableau 13: Taille du corpus selon le Score de Confiance d'openFace

2.1.3. Modèles prédictifs

Afin d'effectuer la tâche de classification, nous avons opté pour des SVM (Séparateurs à vastes marges) et cela pour plusieurs raisons :

- Leur facilité d'utilisation
- Leur rapidité d'exécution
- Leur efficacité à généraliser et leur précision en classification

Les SVM nous ont permis avec une grande rapidité d'obtenir des résultats et d'orienter la recherche sur la détection de contexte.

L'implantation s'est faite via la libraire Scikit-Learn pour python [38]. Cette librairie contient plusieurs modules de machine learning dont plusieurs concernant les SVM. Les expérimentations ont été faites avec le module SVC car notre ensemble de données reste petit et nous cherchons à classifier ces données.

L'utilisation d'un noyau linéaire à fin de classification limite l'optimisation des hyper-paramètres. Dans notre cas, seul l'hyper-paramètre C a été optimisé sur une échelle logarithmique allant de 1-e6 à 1.

Figure 10: UAR (axe Y) Moyen des modèles prédictifs par ensemble de données et complexité (axe X)

Chaque SVM est entraîné sur une partition d'Entraînement puis optimisé sur une partition Validation. Le modèle est ensuite ré-entraîné sur la partition d'Entraînement en prenant en compte les meilleurs hyper-paramètres pour la classification, puis le modèle est évalué sur la partition Test. Cette opération est répétée cinq fois puis l'on calcule la moyenne des résultats sur les cinq itérations. De cette manière on obtient la performance moyenne d'un modèle avec les meilleurs hyper-paramètres. Enfin, l'on calcule la moyenne et l'écart-type de ces modèles sur l'ensemble des jeux de données utilisées (les jeux de données dont la précision d'openFace est supérieure à 85, 90, 95, 96). Cette dernière mesure nous permet de mesurer la fiabilité et la précision d'un modèle sur plusieurs jeux de données.

2.2. Résultats des modèles prédictifs

Une fois l'ensemble des modèles entraînés et optimisés, les résultats sur les partitions Test ont été consignés afin de pouvoir les comparer plus facilement.

2.2.1. Résultats de la classification binaire SAS/VAS

La première étape de détection de contexte concerne les scénarios d'interaction SAS et VAS. Pour rappel, il s'agit d'identifier si le sujet s'adresse à lui-même ou s'il est en interaction avec l'assistant virtuel.

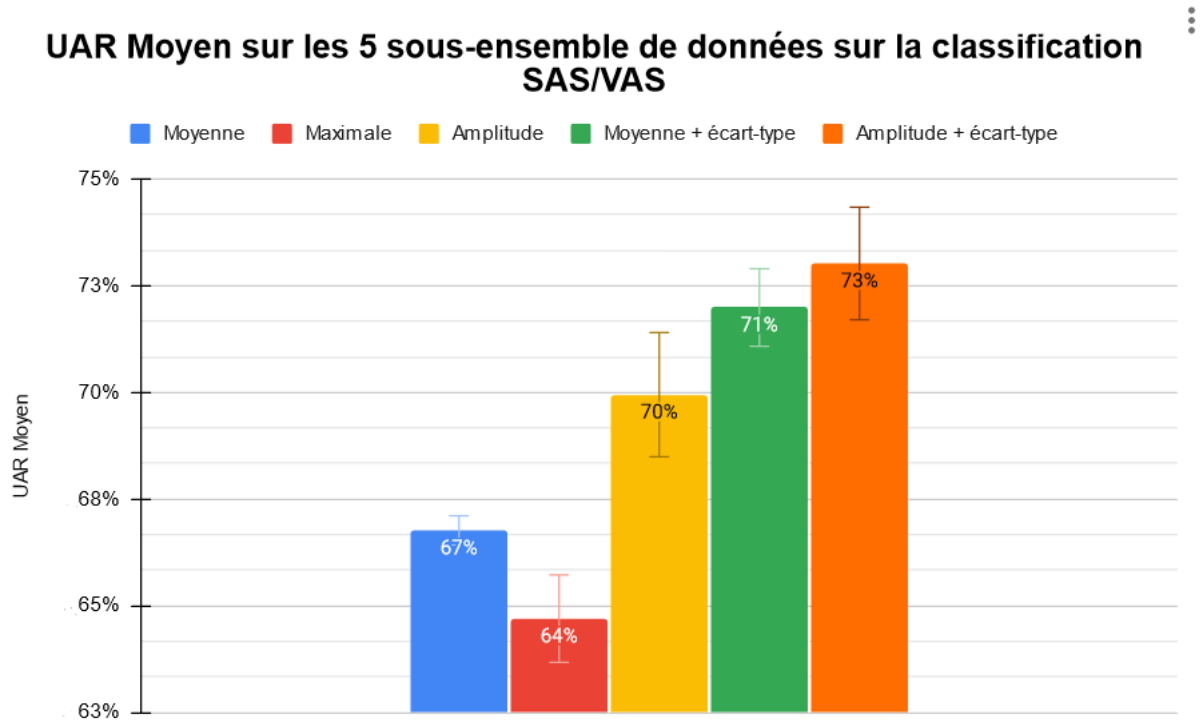

Figure 11: Récapitulatif des performances moyennes des modèles prédictifs

La figure 11 synthétise la performance des différents modèles prédictifs en fonction de la caractéristique extraite. Également, l'écart-type donne un aperçu de la performance d'un modèle sur les différents ensembles de données filtrés. Parmi toutes les caractéristiques, on observe que la moyenne est la plus stable (l'écart-type y est particulièrement faible) et l'amplitude la plus efficace en moyenne. Également, l'ajout de l'écart-type augmente la performance des modèles. L'amplitude et l'écart-type semblent être les deux caractéristiques les plus efficientes pour prédire ces deux premiers contextes avec un taux d'UAR de 73 %. Enfin, comme les statistiques préliminaires le soulignaient, les valeurs maximales sont bien moins discriminantes pour le SVM. On peut émettre à partir de là l'hypothèse que la valeur maximale est une piste moins intéressante pour la détection de contexte.

2.2.2. Résultats de la classification binaire des SI de niveau 2

D’après les conclusions des statistiques préliminaires, on peut s’attendre à des résultats nettement inférieurs quant à la prédiction des scénarios d’interaction de second niveau. Les tests statistiques mettaient en évidence la difficultés à prouver qu’une FAU soit propre à un label.

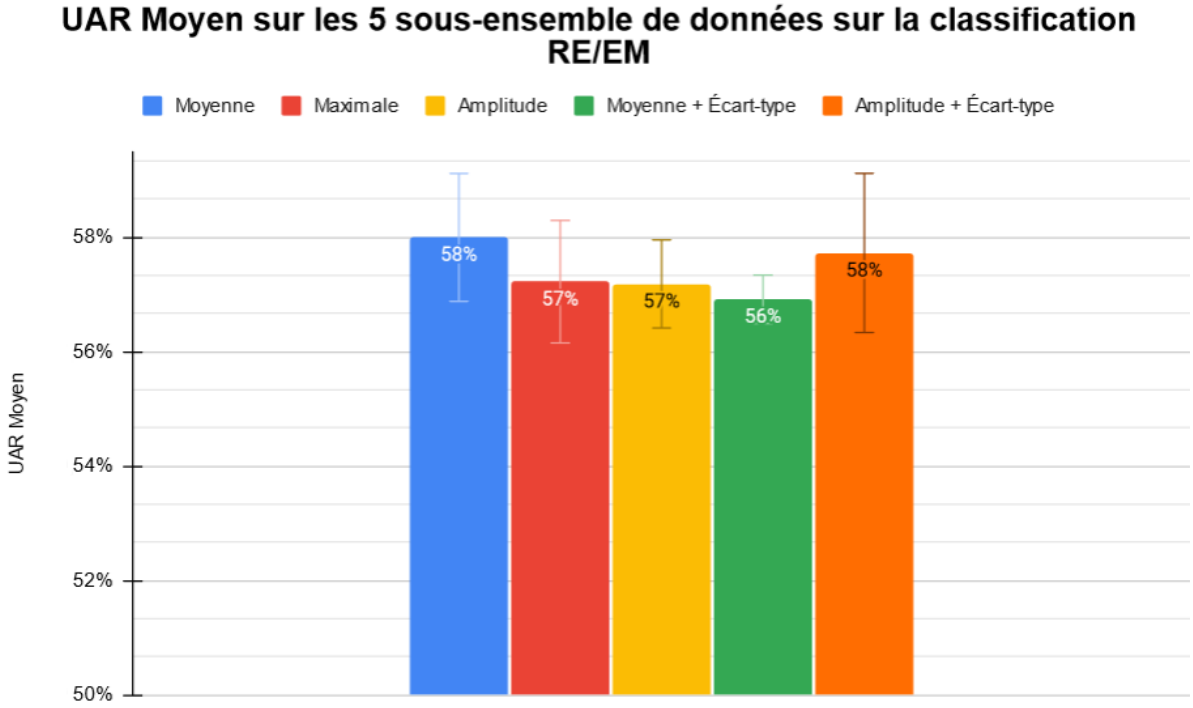

Figure 12: Récapitulatif des performances des modèles prédictifs sur les différents jeux de données. Tâche de classification RE/EM

La figure 12 rend compte de l’UAR moyen des différents modèles prédictifs. Or tous les modèles ont un taux de prédiction entre 56 et 57 %. Les résultats sont donc particulièrement homogènes. De même, on observe une plus grande dispersion des taux de prédiction, notamment pour les valeurs maximales et l’amplitude associées à l’écart-type. Ajoutons que les résultats sont similaires lors de la détection des labels RA et PR qui appartiennent au contexte VAS. Ces résultats mettent en lumière l’extrême difficulté des SVM à prédire un contexte plus précis. Cette baisse de près de 10 % en moyenne peut s’expliquer par le fait que les FAUs peuvent être propre à plusieurs contexte et qu’il semble que les patterns pour les scénarios d’interaction de niveau 2 soient peu apparents ou relativement similaires. Néanmoins, les résultats demeurent au-dessus du score chance (50%), les SVM assurent donc dans une moindre mesure la détection de contexte.

Lors de la multi-classification, on observe également des résultats particulièrement homogènes.

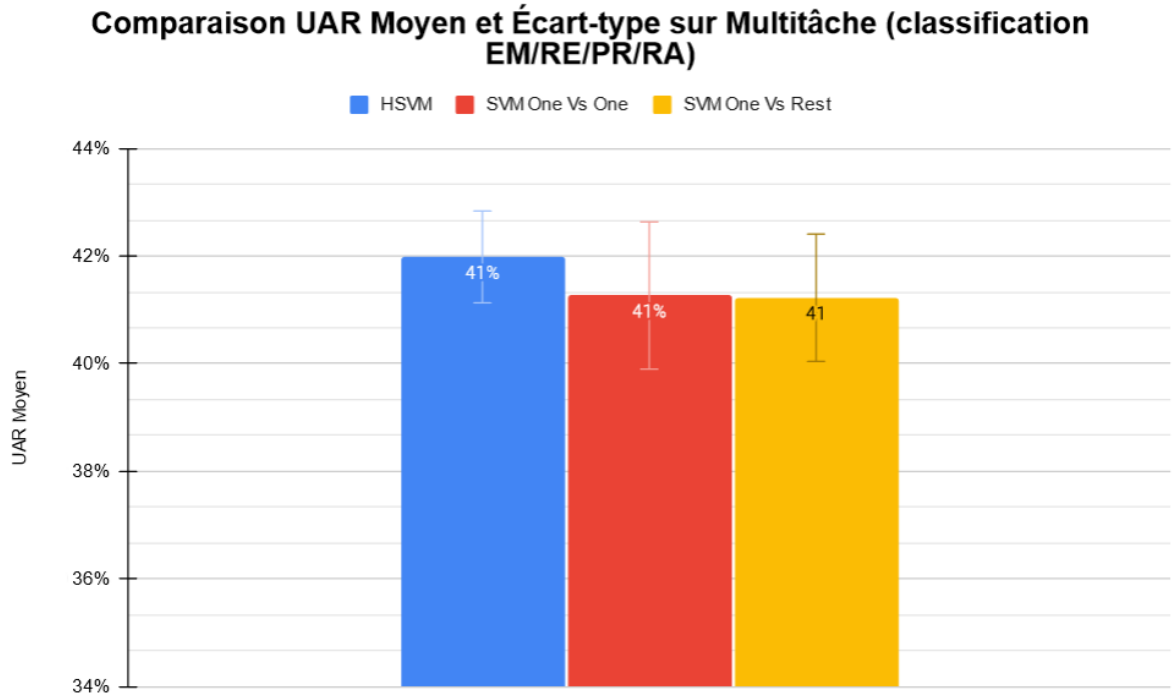

Figure 13: Performance moyenne des SVM multi-tâche sur la classification des labels RE/EM/RA/PR. UAR est la moyenne des modèles sur les différents jeux de données.

Pour cette expérimentation, trois types de modèles ont été utilisés :

- Un modèle HSVM : le modèle HSVM (Séparateur à vaste marge hiérarchique) se compose de trois classifieurs (un pour les labels SAS/VAS, un pour les labels RE/EM et un dernier pour les labels RA/PR). Le premier classifieur prédit les labels SAS et VAS pour un ensemble de données, puis en fonction du résultat de la prédiction, on fait appel au second ou troisième classifieur.
- Un modèle SVC dont la stratégie de décision est *One vs One* (le classifieur simplifie la multi-classification en plusieurs problèmes de classification bianires. Pour une classe C^1 donnée, il va créer des tâches de classification binaire pour la classe C^1 vs C^2, C^1 vs C^n ...)
- Un modèle SVC dont la stratégie de décision est *One vs Rest* (tout comme pour la stratégie One vs One, le classifieur simplifie la multi-classification en tâche de classification binaire. Là différence étant que pour chaque tâche de

classification il décide entre une classe C^1 et toutes les autres classes, par exemple : C^1 vs [C^2, C^3, C^n]

Les HSVM permettent une plus grande robustesse sur l'ensemble des jeux de données. Si les résultats sont dans l'ensemble inférieur aux expériences effectuées jusqu'alors, il faut concevoir que tous nos modèles performent au-delà du score chance (de 25%). La modélisation permet donc de souligner qu'il est possible de reconnaître ces contextes. Néanmoins, ces résultats peuvent très certainement être optimisés.

2.2.3. Détection de contexte

Dernière expérimentation tentée, nous cherchions à savoir si un modèle prédictif était capable de reconnaître un contexte qui nécessite l'intervention de l'assistant virtuel d'un contexte qui ne nécessite pas l'intervention de l'assistant virtuel.

Pour ce faire les quatre labels que nous utilisions jusqu'à présent ont été regroupé sous deux annotations distinctes :

- Une première annotation **I** pour Intervention. Dans ce groupe, l'on retrouve le scénario d'interaction EM ainsi que le groupe de scénario d'interaction PR.
- Une seconde annotation notée **NI** pour Non-Intervention. Dans ce groupe, l'on retrouve le scénario d'interaction RE ainsi que le groupe de scénario d'interaction RA. Précisons que le groupe d'interaction RA contient deux labels, un premier NU qui nécessite l'intervention de l'assistant virtuel, et un second UN qui ne nécessite pas l'intervention de l'assistant virtuel. Le second étant le label le plus fréquent, il a été décidé que le groupe RA serait généralisé en tant que groupe de scénario d'interaction ne nécessitant pas d'intervention.

À l'instar des expérimentations précédentes, les résultats présentés correspondent à la moyenne des prédictions sur 5 partitions Test. On utilise ici le Rappel afin d'observer uniquement les cas pertinents pour le modèle.

Comme on peut le constater sur la Figure 14, les résultats sont relativement bons. Comme on peut s'y attendre, les SVM ont plusieurs difficultés à généraliser sur les données présentées. Le modèle utilisé ici est un modèle Linéaire qui utilise pour caractéristique

l'Amplitude et l'écart-type afin de classifier les données. C'est le modèle qui s'est révélé le plus performant jusqu'à présent.

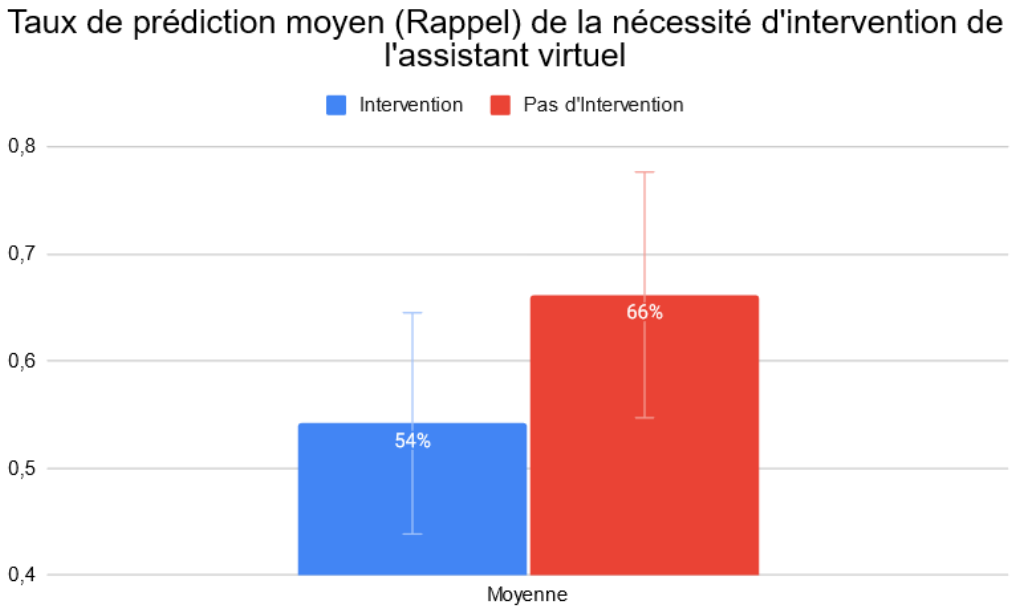

Figure 14: Taux de détection moyen du contexte en fonction de la nécessité d'intervention

Un des problèmes majeurs de la détection d'intervention dans le cas présent est la répartition des labels. Le groupe **I** contient à la fois le label EM et PR, quant au groupe **NI**, il est composé des labels RE et RA. Chaque groupe contient donc un label appartenant au groupe SAS et un label appartenant au groupe VAS. Ajoutons à cela que ces deux groupes étaient particulièrement bien discriminés par nos SVM. Or, plus l'on tâchait d'être précis dans la classification, plus les résultats baissaient. Ici, les SVM classifient difficilement des groupes de scénarios d'interaction qui couvre d'un point de vue des caractéristiques plusieurs contextes différents et variés. À ce niveau-là, il serait judicieux à l'avenir de ne pas utiliser que les FAUs pour détecter la nécessité d'intervention de l'Assistant Virtuel.

CONCLUSIONS

À travers les articles de Scherer, Chetouani ou Pelachaud, l'on comprend que le contexte dans le cadre d'interaction (humain – humain ou humain – machine) est un facteur important dans l'étude des signaux sociaux. L'étude effectuée dans ce mémoire cherche modestement à contribuer à ce pan de la recherche en affective computing.

Ce mémoire a permis de mettre en avant la difficulté d'étudier le contexte. THERADIA propose sa propre grille d'observation et d'interprétation du contexte (les scénarios d'interaction). Néanmoins, cette grille semble très tôt montrer des limites (contextes qui se chevauchent) et pointer du doigt une des difficultés majeurs des études sur le contexte : la variété de contexte et de réaction à un contexte.

L'analyse statistique des FAUs a mis en lumière la capacité à discriminer les deux macro-groupes de notre taxonomie des scénarios d'interaction. Néanmoins, plus l'on cherchait à observer dans le détail la formation des FAUs, plus il apparaissait difficile de les discriminer. Également l'analyse statistique a pu souligner deux caractéristiques principales des FAUS :

- D'une part une FAU peut se retrouver au sein de plusieurs contextes. Ce qui signifie qu'une FAU intense n'est pas nécessairement discriminante.
- Les FAUs forment des patterns. En plus de ces patterns, il serait intéressant d'effectuer une analyse approfondie de l'intensité de chaque FAU au sein des patterns pour permettre de les identifier et les classer.

Néanmoins, les patterns des FAUs peuvent être présent également dans plusieurs contexte. On en déduit donc que les expressions faciales peuvent se montrer particulièrement ambiguës.

Le résultat des analyses statistiques a été corroboré par la modélisation des scénarios d'interaction avec des modèles prédictifs. Plusieurs caractéristiques ont été extraites à partir des FAUs afin d'obtenir différents résultats à comparer. Il est très tôt ressorti que la moyenne était une caractéristique particulièrement stable mais bien moins performante en général que l'amplitude. Qui plus est, l'ajout de l'écart-type à d'autres caractéristiques permet d'obtenir

une plus grande efficacité. Cela est dû fait que les SVM ne prennent pas en compte la séquentialité lors de la discrimination des données. Or l'amplitude et l'écart-type donnent un aperçu de la distribution des valeurs lors d'une trame.

Néanmoins, si les résultats sont satisfaisants lors d'une classification binaire sur les deux macro-groupes (SAS/VAS), les SVM peinent davantage à généraliser sur les autres labels. Cela est principalement dû à deux phénomènes :

- D'une part il y a moitié moins de données lorsque l'on cherche à discriminer les labels au sein du groupe SAS ou VAS
- D'autres part les scénarios d'interactions à l'intérieur des groupes SAS et VAS partagent beaucoup de patterns d'activation des FAUs. C'est ce qui rend particulièrement difficile la classification.

Enfin, l'on a pu observer des résultats également médiocres sur la détection de deux types de contexte différents (intervention, non-intervention). Le problème est dû aux valeurs des FAUs à l'intérieur des deux jeux de données, qui sont particulièrement similaires.

Il faut cependant rester optimiste puisque cette étude a pu mettre en évidence de bonnes capacités à reconnaître certains contextes à partir d'un faible nombre de caractéristiques. Cela laisse entrevoir des possibilités quant à l'amélioration du système de prédiction contextuel en approfondissant la multi-modalité de nos données (orientation du regard, posture, tf-idf, MFCC). De plus, les statistiques et les modèles prédictifs utilisés soulignent *a minima* l'importance des informations véhiculées par les FAUs lors d'interaction en dyade.

Bibliographie

[1] Weber, M. (1978). *Economie et Société*

[2] Turner, JH. (1988). *A Theory of social interaction*

[3] Coker, D. A., and Burgoon, J. (1987). The nature of conversational involvement and nonverbal encoding patterns. *Hum. Commun. Res.* . 13. pp.463–494

[4] Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, et al. (2020). Engagement in Human-Agent Interaction: An Overview. *Frontiers in Robotics and AI,* Frontiers Media S.A.,.7. pp.92.

[5] Argyle.M. *The Psychology of interpersonnal behaviour*. Harmondsworth. 1967.

[6] Ekman, P., Friesen, W.F. (1969). The repertoire of nonverbal behavioral categories – origins, usage, and coding. *Semiotica.* Vol. 1, pp.49–98.

[7] Pantic, M., Nijholt, A., Pentland, A. and Huanag, T.S. (2008). Human-Centred Intelligent Human– Computer Interaction (HCI2): how far are we from attaining it?. *Int. J. Autonomous and Adaptive Communications Systems.* Vol. 1. No. 2. pp.168–187.

[8] Juslin, P. N., & Scherer, K. R. (2005). Vocal expression of affect. In J. A. Harrigan, R. Rosenthal, & K. R. Scherer (Eds.). *The new handbook of methods in nonverbal behavior research.* Oxford University Press. pp. 65–135.

[9] R. Cowie *et al*. (2001) .Emotion recognition in human-computer interaction. In *IEEE Signal Processing Magazine. V*ol. 18. no. 1. pp.32–80. doi: 10.1109/79.911197.

[10] Ekman, P., & Friesen, W. V. (1978). Facial action coding system. *CA: Consulting Psychologists Press.*

[11] Gottman, J., Levenson, R., Woodin, E. (2001). Facial Expressions During Marital Conflict, *Journal of Family Communication.* 1:1. pp–37-57.

[12] André, E., Pelachaud, C. (2010). Interacting with Embodied Conversational Agents. In: Chen, F., Jokinen, K. (eds) *Speech Technology.* Springer, New York, NY.

[13] McNeill, D. (1992) Hand and Mind: What Gestures Reveal about Thought. University of Chicago.

[14] Knapp, M., Hall, J. (1997). Nonverbal communication in human interaction

[15] Scheflen. A.E. (1964). The Signifiance of posture in communication systems. *Psychiatry*. 27:316-31.

[16] Coppin G., Sander D., (2010) Théories et concepts contemporains en psychologie de l'émotion, E3Lab, Section de psychologie, FPSE, Université de Genève.

[17] Mahmud, S., Lin, X., Kim, J. -H., (2020). Interface for Human Machine Interaction for assistant devices: A Review. *2020 10th Annual Computing and Communication Workshop and Conference (CCWC).* pp.768–773.

[18] Schneeberger, T.,Scholtes, M., Hilpert, B., Langer, M., Gebhard, P., (2019). Can Social Agents elicit Shame as Humans do?. *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII).* pp.164-170.

[19] Vinciarelli, A., Esposito, A., André, E. *et al.* (2015). Open Challenges in Modelling, Analysis and Synthesis of Human Behaviour in Human–Human and Human–Machine Interactions. *Cogn Comput* **7**, pp.397–413.

[20] Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., & André, E. (2013). The social signal interpretation (SSI) framework: multimodal signal processing and recognition in real-time. In *Proceedings of the 21st ACM international conference on Multimedia.* pp.831–834.

[21] Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., ... & Pantic, M. (2016). Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge.* pp.3–10

[22] Vinciarelli, A., Pantic M., and Bourlard H., (2009). Social signal processing: Survey of an emerging domain. Image and vision computing 27.12. pp.1743–1759.

[23] Gebhard, P., Schneeberger, T., Baur, T., & André, E. (2018). Marssi: Model of appraisal, regulation, and social signal interpretation.

[24] Khan, R., & Sharif, O. (2017). A literature review on emotion recognition using various methods. *Global Journal of Computer Science and Technology.*

[25] Schuller, B., Weninger, F., Zhang, Y., Ringeval, F., Batliner, A., Steidl, S., ... & Mortillaro, M. (2019). Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge. *Computer Speech & Language.* 53. pp.156–180.

[26] Kim, S., Filippone, M., Valente, F., Vinciarelli, A., (2012). Predicting the conflict level in television political debates: an approach based on crowd-sourcing, nonverbal communication and gaussian processes. In: *Proceedings of ACM Multimedia. ACM*, Nara, Japan, pp.793–796.

[27] Banziger, T., Mortillaro, M., Scherer, K.R., (2012). Introducing the Geneva multimodal expression corpus for experimental research on emotion perception. *Emotion,* 12, pp.1161–1179.

[28] Ringeval, F., Demouy, J., Szaszak, G., Chetouani, M., Robel, L., Xavier, J., Cohen, D., Plaza, M., (2011). Automatic intonation recognition for the prosodic assessment of language impaired children. *IEEE Trans. Audio Speech Lang. Process.* 19, pp.1328–1342.

[29] Kim, J., Truong, K.P., Charisi, V., Zaga, C., Lohse, M., Heylen, D.K., Evers, V. (2015). Vocal turn-taking patterns in groups of children performing collaborative tasks: an exploratory study. *INTERSPEECH*.

[30] Kraaij, W., Hain, T., Lincoln, M., Post, W. (2005). The AMI meeting corpus.

[31] Tarpin-Bernard, F., Fruitet, J., Vigne, J. P., Constant, P., Chainay, H., Koenig, O., ... & Ghenassia, D. (2021). THERADIA: Digital Therapies Augmented by Artificial Intelligence. In *International Conference on Applied Human Factors and Ergonomics.* Springer. Cham. pp. 478–485.

[32] Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation.* 41(3). pp.273–287.

[33] Saint-Georges, C., Mahdhaoui, A., Chetouani, M., Cassel, R. S., Laznik, M. C., Apicella, F., ... & Cohen, D. (2011). Do parents recognize autistic deviant behavior long

before diagnosis? Taking into account interaction using computational methods. *PloS one*, 6(7).

[34] ELAN (Version 6.5) [Computer software]. (2023). Nijmegen: *Max Planck Institute for Psycholinguistics*, The Language Archive. Lien URL : https://archive.mpi.nl/tla/elan

[35] Baltrušaitis, T. , Zadeh, A., Lim, Y. C., Morency, L-P., *(*2018) openFace 2.0: Facial Behavior Analysis, *IEEE International Conference on Automatic Face and Gesture Recognition.*

[36] Bellard, F., (2023). FFmpeg tool (Version 6.0) [Computer Software]. Available from http://ffmpeg.org/

[37] Farnsworth, B. (2022, Mai 12), *Facial Action Coding System (FACS) – A Visual Guidebook*, Imotions. https://imotions.com/blog/learning/research-fundamentals/facial-action-coding-system/

[38] Pedregosa *et al.* (2011). Scikit-learn : Machine Learning in Python. *JMLR 12*. pp. 2825-2830.

ANNEXES

I. Liste des abréviations utilisées

UA/AU/FAU : Unité d'Action, Action Unit, Facial Action Unit

SI /IS : Scénario d'interaction, Interaction scenario

SAS : Self-Adressed Speech (Parole auto-adressé)

VAS : Virtual Assistant Adressed Speech (Parole adressé à l'assistant virtuel)

RE : Reflectional (Réflective)

EM : Emotional (Émotionnelle)

GU : Giving-Up (Abandon)

RA : Rétro-active

NU : Not Understanding (Incompréhension)

UN : Understanding (Compréhension)

PR : Propositionnal (Propositionnel)

APP : Approval (Approbation)

OT : Others (Autres)

II. Guide de transcription et de segmentation THERADIA

Contexte

Ce document a pour objectif de préciser les modalités de réalisation des tâches de transcription et de segmentation des enregistrements collectés dans le projet THERADIA. Pour rappel, un ensemble de sujets, jeunes et âgés, ont participé à des sessions d'entraînement cognitif en étant accompagné par un assistant virtuel – piloté par un agent humain en tant que magicien d'Oz. Les flux audiovisuels capturés par notre dispositif ont été enregistrés et l'objectif est de fournir à la fois une segmentation et une transcription de ces enregistrements.

Segmentation en propositions

La segmentation est réalisée sous la forme de propositions, i.e., des suites continues ou discontinues de mots qui portent une même information. Cette information est de nature linguistique et est portée par la sémantique. Autrement dit, la prosodie et les expressions faciales ne servent pas de support temporel pour la segmentation d'une proposition, mais d'éléments contextuels, qui sont intégrés à la proposition s'ils occurrent durant son expression ou à proximité immédiate. Par exemple, si un sujet affiche une expression faciale notable juste avant de s'exprimer, cette expression doit être contenue dans la proposition. Cela est également le cas si l'expression faciale intervient après l'expression verbale, ou si les deux sont présentes, ces dernières doivent être conservées et intégrées à la proposition.

La raison pour laquelle nous souhaitons effectuer une segmentation sous la forme de propositions vient du fait que nous devons pouvoir interpréter le sens du message verbal, et le contextualiser par les informations non-verbale associées. Il est donc nécessaire de découper le flux de parole en segments qui renvoient à une seule information, permettant de simplifier la compréhension du langage naturel qui est réalisée par des méthodes automatiques. Ainsi, si une personne enchaîne rapidement plusieurs groupes de souffle qui renvoient à la même information, ils constituent une seule proposition. Si ces groupes de souffle renvoient chacun

à un type d'information différent, ils constituent chacun une proposition. Par exemple, le flux de parole continu dont la transcription est « alors heu [pause] je me suis plutôt trompé sur cet exercice [pause] mais je pense avoir plutôt bien réussi » constitue une seule proposition puisque la personne donne son impression sur un même sujet, à savoir sa performance sur l'exercice réalisé.

Expressions prosodiques, faciales, et posturales

Pour rappel, le support verbal prime sur le support non-verbal. Autrement dit, une variation importante de charge affective dans un flux continu de parole ne déclenche pas la définition de deux propositions distinctes, à moins que la sémantique renvoie à des informations différentes. Les expressions faciales notables situées à distance des propositions doivent être segmentées sous la forme de proposition non-verbale à l'aide du diacritique <nv>.

Contextualisation des propositions

Dans le cas où une expression faciale est produite à proximité d'une proposition, cette dernière doit être incluse de sorte de fournir, dans la mesure du possible, un peu de contexte avant et après la proposition. Le positionnement des marqueurs temporels de début et fin de proposition ne doivent donc pas être situés le plus près possible du premier et dernier phonème prononcé, mais situés à distance de façon à fournir un espace d'observation suffisant pour pouvoir analyser la proposition. Comme ces séquences seront annotées par la suite en temps-continu, les annotateurs devront en effet avoir à disposition un contexte suffisant pour évaluer la proposition. Notons qu'il est préférable d'avoir une durée du contexte avant la proposition plus grande que celle suivant la proposition, et que la durée du contexte doit se limiter à quelques secondes.

Conventions de transcription

Les conventions de transcription utilisées dans ce document suivent en grande partie celles du projet ESLO avec les deux principes de base suivant : le respect de l'orthographe et de ce qui a été dit, notamment en termes de structures grammaticales.

Dictionnaire

Le dictionnaire de référence est « Le Petit Robert », accessible en ligne à l'adresse suivante : https://dictionnaire.lerobert.com/. Les mots non attestés dans le dictionnaire sont précédés du symbole et commercial (il était &relou).

Point d'interrogation

Le point d'interrogation est le seul signe de ponctuation utilisé et doit toujours être précédé d'un espace.

Guillemets, majuscules et apostrophes

L'usage des guillemets est proscrit tout comme pour les majuscules (sauf pour les noms propres), et les apostrophes ; y'a > y a, hormis pour les cas d'usage orthographique ; e.g., « c'est qu'y a eu un changement ».

Élisions

Il faut conserver dans la graphie le schwa s'il garde son contenu mélodique sans être assimilé à « euh », en ne marquant pas l'élision par l'apostrophe ; e.g., « parce que on ».

Troncations

Lorsqu'un mot est tronqué, on utilise le trait d'union pour marquer la troncation ; e.g., « c'était vrai- vraiment pas facile ».

Rétablissements

Lorsque l'expression verbale est déformée et qu'elle appartient au lexique, elle est conservée. En revanche, une déformation qui ne rentre pas dans le lexique est rétablie ; e.g., oblette > omelette, rénuméré > rémunéré. On ne rétablit pas le « il » de « il y a » s'il n'est pas prononcé. Les élisions erronées ne sont pas rétablies même si elles ne sont pas correctes sur le plan orthographique ; e.g., « c'est lui qu'était malade ». Les aphérèses sont rétablies car elles nuisent à la compréhension du langage ; e.g., -tendez > attendez ; fin > enfin.

Onomatopées et interjections

L'orthographe à utiliser pour les onomatopées et interjections est la suivante :

ah aïe ; bah beh ben bof bouh boum ; clac ; euh ; hé hein hop hou ; la ; oh ouais ouf ouille oulala oups ; mouais miam moui ; pff ; roh ; ta ; tac ; zut ;

Diacritiques

La liste des diacritiques possibles est la suivante :

<?> A utiliser lors d'un doute sur la segmentation/transcription, ces cas sont discutés chaque semaine de façon collégiale

<di> A placer juste avant la première proposition à destination de l'assistant virtuel (début interaction)

<fi> A placer juste après la dernière proposition à destination de l'assistant virtuel (fin interaction)

<nv> A utiliser pour définir le contenu des proposition non-verbales situées à distance des propositions verbales

A utiliser à l'intérieur d'une proposition lorsqu'il y a présence d'une expression non-verbale (affective ou autre) marquée

Interruptions séance

Il arrive parfois qu'au cours d'une séance, un membre du personnel universitaire doive intervenir. Lorsque cela est le cas, la séance est interrompue, et l'on ne reprend la segmentation et la transcription que lorsque la séance reprend.

Taxonomie SI

L'étiquetage des différents scénarios d'interaction se fait en fonction de la situation d'énonciation.

Si le sujet se parle à lui-même (parole auto-adressée), ou s'exprime à l'assistant virtuel (parole adressée), plusieurs types de scénarios d'interactions sont possibles ; soit l'expression peut être réflective, chargée en émotion ou communiquer une volonté d'abandon, soit elle peut être réalisée comme élément de rétroaction pour manifester de la compréhension ou de l'incompréhension, ou comme un espace de parole ouvert dont l'objectif peut varier selon le contexte (besoin d'être rassuré ou de simplement discuter) :

- **SAS (Self-Addressed Speech / Parole Auto-Adressée)**

- RE (Reflectional / Réflective) : le sujet n'est pas en interaction avec l'assistant virtuel et est dans une étape de réalisation d'exercice ou d'entraînement. Il exprime sa réflexion ou réagit à des évènements (ex : découverte de l'exercice) sans charge émotionnelle particulière (répète la consigne, lit à voix basse, s'auto-juge ou s'auto-questionne sans avis tranché en exprimant un débat interne de façon neutre).
- EM (Emotional / Émotionnelle) : le sujet n'est pas en interaction avec l'assistant virtuel et est dans une étape de réalisation d'exercice ou d'entraînement. Il exprime sa réflexion ou réagit à un évènement avec une charge émotionnelle (ne comprends pas le principe de l'exercice ou sa mécanique, n'a pas réussi un exercice une première fois ou seconde fois, critique leur résultats ou le fonctionnement de l'exercice, ils/elles sourient, hochent la tête, froncent les sourcils, baillent, émettent des interjections / jurons, …).
- GU (Giving-Up / Abandon) : le sujet n'est pas en interaction avec l'assistant virtuel et est dans une étape de réalisation d'exercice ou d'entraînement. Il exprime sa volonté ou son souhait d'abandonner l'exercice en cours. Les sujets peuvent manifester cette volonté d'abandonner l'exercice soit de façon explicite en le verbalisant clairement, soit ils l'expriment de façon indirecte en manifestant des signes de lassitude ou de fatigue extrême, le ton s'intensifie alors nettement, des injures peuvent être prononcées, les soupirs sont nombreux et profonds.

- **VAS (Virtual assistant Addressed Speech / Parole Adressée à l'Assistant Virtuel) :**
  - NU (Not-understanding / Incompréhension) : le sujet n'a pas compris la consigne ou entendu l'assistant virtuel, ou la mécanique de l'exercice en cours, ou les raisons possibles d'échecs, (le sujet se rapproche de l'écran, plisse les paupières, demande à l'assistant virtuel de répéter la question, la consigne, répète plusieurs fois une même phrase où il manifeste une forme de détresse, demande de recommencer un exercice, montre une forme de déception ou d'agacement, ou de surprise par rapport à ses résultats).

- UN (Understanding / Compréhension) : le sujet manifeste sa compréhension des consignes données par l'assistant virtuel (hoche la tête verticalement, ferme les paupières de façon prolongée, exprime des remerciements ou des formes d'acquiescement et marquant la compréhension de façon générale ; "d'accord","hum","ok", …).
- APP (Approval / Approbation) : le sujet demande à être rassuré ou approuvé par l'assistant virtuel (il interroge l'assistant virtuel sur ses propres réponses, lorsqu'il fait une fausse manipulation, ou sur ses performances).
- OT (Others / Autres) : tous les autres cas non spécifiés, typiquement lorsque l'assistant virtuel pose une question au sujet ou lorsque ce dernier s'exprime librement ou manifeste son souhait de prendre la parole ; il n'est pas nécessaire de mettre d'étiquette dans ce cas de figure et le champ peut rester vide.

III .Table des tableaux et figures

1. Table des Figures

Index des figures

2. Table des Tableaux

Index des tableaux

IV. Table des matières

Table des matières

Résumé :

L’objectif de ce mémoire est d’exploiter différents types de contexte d’interaction entre un humain et un assistant virtuel dans le but d’améliorer les performances d’un système de reconnaissance d’émotion. Nous allons pour cela travailler sur le corpus THERADIA-WoZ qui contient des enregistrements d’interactions entre des sujets seniors et une assistante virtuelle lors de la réalisation d’exercice de remédiation cognitive. Ces données ont été segmentées en termes de proposition verbale ou non-verbale, transcrites, et annotées selon des dimensions affectives et interactionnelles. L’objectif de cette étude est de vérifier s’il est possible de détecter dans un premier temps les différents types de contexte interactionnels à partir des enregistrements audiovisuels, d’étudier ensuite sur le plan statistique les dépendances éventuelles entre dimensions affectives et interactionnelles, et enfin de détecter lorsque l’assistant virtuel doit intervenir dans le cadre d’une interaction.

Mots-clefs : Interaction, contexte, machine learning, assistant virtuel, émotions

Abstract :

The aim of this thesis is to exploit different types of interaction context between a human and a virtual assistant in order to improve the performance of an emotion recognition system. To this end, we will work on the THERADIA-WoZ corpus, which contains recordings of interactions between senior subjects and a virtual assistant during cognitive remediation exercises. These data were segmented in terms of verbal and non-verbal propositions, transcribed and annotated according to affective and interactional dimensions. The aim of this study is to ascertain whether it is possible to detect the different types of interactional context from the audiovisual recordings, then to study statistically the possible dependencies between affective and interactional dimensions, and finally to detect when the virtual assistant should intervene as part of an interaction.

Keywords : Interaction, context, machine learning, virtual assistant, emotions