Leveraging the dynamics of non-verbal behaviors : modeling social attitude and engagement in human-agent interaction

Soumia Dermouche

EDITE - ED 130

THÈSE DE DOCTORAT DE
Sorbonne Université

Spécialité

Informatique

École doctorale Informatique, Télécommunications et Électronique (Paris)
présentée et soutenue publiquement par

Soumia DERMOUCHE

le 27 juin 2019

Leveraging the Dynamics of Non-Verbal Behaviors: Modeling Social Attitude and Engagement in Human-Agent Interaction

Directrice de thèse: **Catherine PELACHAUD**

devant le jury composé de :

M. Thierry DUTOIT, Professeur, NUMEDIART, Université de Mons	Rapporteur
M. Alexandre PAUCHET, Maître de conférences, LITIS, INSA Rouen	Rapporteur
M. Mohamed CHETOUANI, Professeur, ISIR, Sorbonne université	Examinateur
Mme Elisabeth ANDRÉ Professeur, Augsburg University	Examinateur
Mme Magalie OCHS, Maître de conférences, LSIS, Aix-Marseille Université	Examinateur
Mme Catherine PELACHAUD, Directrice de recherche, CNRS-ISIR	Directrice de thèse

À mes chèrs enfants

Remerciement

La réalisation de cette thèse n'aurait pas abouti sans la présence, l'aide et la contribution de de plusieurs personnes. Premièrement, je voudrais remercier ma Directrice de thèse Catherine Pelachaud pour m'avoir donné cette opportunité. Grâce à toi j'ai découvert le sens de la persévérance et de l'exigence qui font naître un travail de qualité. Merci infiniment pour ta disponibilité malgré toutes tes responsabilités.

Je remercie également tous les membres de mon jury: Thierry Dutoit, Alexandre Pauchet, Mohamed Chetouani, Elisabeth André et Magalie Ochs. Merci pour le temps investi à évaluer mon travail ainsi que pour vos questions et remarques pertinentes.

Mes collègues de la formidable Greta Team ont énormément contribué à ma thèse en m'offrant de l'écoute, une bonne ambiance et de l'aide. Je commence par les anciens qui étaient déjà là quand je suis arrivée : Mathieu (merci pour le bagage de démarrage : corpus, annotations...), Nesrine, Florian, Brian, Nadine, Caroline, Irina, Thomas et Valentin, je vous remercie pour les agréables moments et les riches discussions partagés avec moi. Merci à nos ingénieurs André-Marie, Brice et Philippe sans oublier le génie post-doc Angelo qui m'ont aidé à trouver des solutions à tous mes problèmes VIB. Je remercie et souhaite une bonne continuation pour les nouveaux arrivants : Léo, Reshma, Fajrian, Sooraj, Fabien et Tanvi.

Je souhaite profondément remercier ma copine de galère Beatrice qui m'a beaucoup aidée et qui a partagé avec moi beaucoup de difficultés que nous avons rencontrées tout au long de notre thèse : changement de l'école doctorale, prolongation de contrat, inscription en 4ème année, mission d'enseignement... Merci aussi pour ton aide scientifique : les tests statistiques, les FMLs de mon expérience à la cité des sciences... Merci à Chloé et Giovanna pour vos remarques et commentaire très intéressants qui m'ont aidée à mieux présenter mon travail.

Grand merci à mes stagiaires Mohamed, Katia et Mohand-Cherif pour l'annotation de mon corpus. Merci également à Maurizio d'avoir collaboré avec moi pour intégrer EyesWeb dans VIB. Je voudrais remercie Camille et Hélène du Carrefour Numérique de la cité des sciences et de l'industrie qui m'ont accordé une formidable occasion pour évaluer mon modèle à la cité des sciences.

Pour finir, je voudrai remercier ma famille en particulier mes parents et ma sœur pour leur soutien et encouragement malgré la distance qui nous sépare. Merci à mon mari pour ton écoute, soutien, encouragement, patience et relecture! Je finis par remercier mes trésors Sirine, Adam et Ayoub qui représentent la source de mon énergie positive.

Abstract

SOCIAL interaction implies exchange between two or more persons, where they adapt their behaviors to each others. With the growing interest in human-agent interactions, it is desirable to make these interactions natural and human like. In this context, we aimed at enhancing the quality of the interaction between users and Embodied Conversational Agents ECAs by (1) endowing the ECA with the capacity to express social attitudes, such as being friendly or dominant depending its role or relationship with its interaction partners; (2) adapting the agent's behavior according to the user's behavior, hence, the conversation partners influence each others through an interaction loop, thus, enhancing the interaction quality; (3) predicting the user's engagement level and adapting the agent's behavior accordingly, which helps maintaining the user's interest and motivation. We take advantage of the recent advances in machine learning, more specifically, temporal sequence mining and neural networks to model these capacities in the ECA. The first model is used to learn relevant patterns (sequences) of non-verbal signals that best represent attitude variations, and then reproduce them on the agent. The latter is used to encompass the dynamics of non-verbal signals (temporal change) to achieve more accurate prediction of behavior. Two use cases have been explored using the well-known LSTM model: agent's behavior adaptation based on both agent's and user's behavior history, and user's engagement prediction based on his/her own behavior history. The implemented models and algorithms have been validated through a number of perceptive studies, in particular performed in Musée des Sciences et de l'Industrie in Paris, as well as through rigorous quantitative analysis of the obtained results. In addition, the realized models have been integrated into a virtual-agent platform.

Keywords: non-verbal behavior, social attitude, engagement prediction, human-agent interaction, temporal sequence mining, virtual agents

Résumé

DANS le contexte de l'interaction humain-agent, notre objectif était d'améliorer la qualité de l'interaction en: (1) dotant l'agent de la capacité d'exprimer des attitudes sociales telles que la dominance ou l'amicalité ce qui renforcent ses compétences sociales; (2) adaptant le comportement de l'agent selon le comportement de l'utilisateur, par conséquent l'agent et l'utilisateur s'influencent mutuellement par le biais d'une boucle interactive; (3) prédisant le niveau d'engagement de l'utilisateur et adaptant en conséquence le comportement de l'agent, ce qui contribue à maintenir l'intérêt et la motivation de l'utilisateur. Nous nous basons sur les progrès récents dans le domaine de l'apprentissage automatique, plus particulièrement de l'extraction de séquences temporelles et des réseaux de neurones. Le premier est utilisé pour apprendre des séquences pertinentes de signaux non-verbaux qui représentent au mieux les variations d'attitude, puis les reproduire par l'agent. Le second est utilisé pour englober la dynamique des signaux non-verbaux. Deux cas d'utilisation ont été explorés à l'aide du modèle LSTM: l'adaptation du comportement de l'agent en fonction de l'historique de comportement de l'agent et de l'utilisateur; et la prédiction de l'engagement de l'utilisateur basée sur son propre historique de comportement. La pertinence des modèles et des algorithmes implémentés a été validée au moyen de nombreuses études approfondies et d'une évaluation quantitative rigoureuse des résultats obtenus. De plus, les travaux réalisés ont été intégrés dans une plateforme d'agents virtuels.

Mots-clefs : comportement non-verbal, attitude sociale, prédiction d'engagement, interaction humain-agent, extraction de séquence temporelle, agents virtuels

Long Résumé

LES agents conversationnels animés (ACAs) sont des personnages virtuels capables d'interagir de manière autonome avec des humains en imitant leurs comportements naturels. *Yoko* de Toshiba et *Tim* d'Airbus sont des exemples d'agents virtuels animés qui répondent aux questions techniques et commerciales des clients. Ces dernières années, les agents virtuels sont devenus de plus en plus présents dans notre vie quotidienne. Ils peuvent être utilisés pour des applications diverses allant de l'éducation et la formation à la thérapie [Nojavanasghari et al., 2016, Chollet et al., 2017, Nojavanasghari and Hughes, 2017]. Par conséquence, de nombreuses recherches ont été consacrées à l'amélioration de l'interaction humain-agent. Compte tenu de l'intérêt croissant que suscitent les interactions humain-agent, il est souhaitable de rendre ces interactions agréables et plus humains. Dans le cadre de cette thèse, nous visons à améliorer l'expérience d'interaction entre les humains et les agents virtuels. À cette fin, nous développons des modèles computationnels pour doter un ACA de la capacité: (1) d'exprimer différentes attitudes sociales en fonction du contexte de l'interaction, (2) d'adapter le comportement de l'ACA en fonction du comportement de l'utilisateur, (3) de prédire le niveau d'engagement de l'utilisateur durant l'interaction humain-agent. Notre objectif est d'enrichir l'état de l'art avec des modèles et des algorithmes plus adaptés et plus fins.

Les humains expriment différentes attitudes sociales les uns envers les autres en fonction du contexte de l'interaction, qui inclut des facteurs tels que l'interlocuteur, le rôle, la personnalité, etc. Par exemple, une personne peut montrer une sorte de *dominance* dans un contexte professionnel alors qu'elle peut être *amicale* lors de sorties entre amis. La même personne ne se comportera pas de la même manière dans ces différentes circonstances. Elle n'aura pas les mêmes comportements. Elle peut utiliser un langage plus formel au travail, afficher une posture plus droite, sourire moins, alors qu'elle peut rire et faire des gestes plus expressifs avec ses amis et sa famille. Dans ce contexte, nous visons à doter un agent virtuel de la capacité d'exprimer des attitudes sociales en fonction du contexte de l'interaction. Par exemple, l'agent devrait être amical avec un client en répondant à sa question mais plus dominant avec un candidat à une offre d'emploi afin de le former à passer des entretiens d'embauches.

Pour une compréhension plus profonde de l'expression de l'attitude sociale, nous devons d'abord explorer ce qui fait qu'une personne apparaît plus ou moins dominante ou encore plus ou moins amicale. Notre question de recherche est : quels sont les comportements non-verbaux qui déclenchent un changement (variation) dans la perception des attitudes sociales? Une telle analyse devrait s'appuyer sur la dynamique des comportements non-verbaux, qui est très informative pour caractériser et interpréter les attitudes. D'autre part, les humains ont tendance à adapter leur comportement tout au long de

l'interaction en fonction du comportement de leur interlocuteur [Burgoon et al., 2010]. Par exemple, une personne hoche sa tête pour indiquer son accord avec l'interlocuteur ou sourit en réponse au sourire de son interlocuteur. Les agents virtuels doivent prendre en compte le comportement de l'utilisateur afin d'adapter et modifier leur comportement en réponse aux actions et comportements de l'utilisateur. Une telle interaction dynamique aide à maintenir l'engagement de l'utilisateur dans l'interaction. Au cours des dernières années, la modélisation de l'engagement a de plus en plus retenu l'attention de chercheurs grace à son rôle important dans l'interaction humain-agent. L'agent doit pouvoir détecter, en temps réel, le niveau d'engagement de l'utilisateur afin de réagir d'une manière appropriée. Dans ce contexte, notre objectif est de développer un modèle computationnel permettant de prédire le niveau d'engagement de l'utilisateur en temps réel. En nous basant sur des résultats antérieurs, nous utilisons les expressions faciales comme caractéristiques (*features*) prédictives de l'engagement [Allwood and Cerrato, 2003, Castellano et al., 2009c]. De plus, l'engagement ne se mesure pas seulement à partir des signaux simples, mais également à partir de la combinaison de plusieurs signaux apparaissant durant une certaine fenêtre temporelle [Peters et al., 2005, Bickmore et al., 2012]. Ainsi, pour une meilleure prédiction de l'engagement, nous devrions considérer la variation des expressions faciales au fil du temps.

Contexte théorique

Le chapitre 2 présente les bases théoriques, les définitions, la représentation et l'expression des attitudes. Une vue d'ensemble des définitions de l'attitude nous a permis de conclure que les attitudes sont *interpersonnelles*, *multimodales* et *dynamiques*. Dans notre travail, nous nous intéressons aux attitudes interpersonnelles, c'est-à-dire les attitudes exprimées envers une personne, en particulier l'attitude que notre agent virtuel exprimera envers l'utilisateur. Le terme *multimodal* signifie que les attitudes sont exprimées verbalement et non-verbalement. Selon Argyle, les deux modalités contribuent de la même façon à l'expression des attitudes [Argyle, 1988]. Par conséquent, nous nous sommes intéressés à l'expression non-verbale de l'attitude. Enfin, les attitudes ne sont pas statiques et varient dans le temps. Notre objectif est d'englober la dynamique des attitudes en considérant conjointement la séquentialité et la temporalité des signaux non-verbaux. En ce qui concerne la représentation des attitudes, différentes dimensions affectives peuvent être utilisées. Le circumplex interpersonnel (CIP) est la représentation la plus populaire des attitudes dans le domaine des agents virtuels. Le CIP est composé de deux dimensions orthogonales comme illustré sur la Figure 1: amicalité (variant de "hostile" à "amical") et dominance (variant de "soumis" à "dominant"). Les deux dimensions de CIP ont été utilisées pour la première fois par Leary [Leary, 1957] qui souligne que chaque comportement interpersonnel ou social peut être représenté, sur le circumplex, comme une combinaison pondérée de dominance et d'amicalité.

Le CIP est récemment devenu un modèle populaire pour évaluer les dispositions interpersonnelles telles que les problèmes interpersonnels (par exemple, les problèmes liés à l'agression d'autrui) [Alden et al., 1990], la valeur (en quoi des expériences interperson-

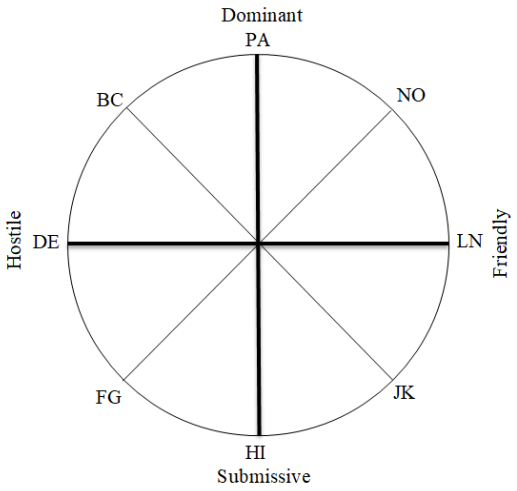


Figure 1 – The interpersonal circumplex (IPC).

nelles, telles que l'expression ouverte, sont-elles importantes pour une personne?) [Locke, 2000], l'efficacité personnelle (actions interpersonnelles qu'une personne croit pouvoir exprimer) et les traits de personnalité [Wiggins, 1995]. La plupart des mesures divisent le CIP en huit octants alphabétiquement libellés dans le sens inverse des aiguilles d'une montre: PA, BC, DE, FG, HI, JK, LM et NO (voir Figure 1). Chaque octant peut être représenté par un ensemble d'adjectifs, par exemple, *dominant* et *assertive* pour l'octant PA. Le *Interpersonal Check List* (ICL), *Interpersonal Adjective Scales* (IAS) et *Inventory of Interpersonal Problems* (IIP) sont des exemples de mesures interpersonnelles. Dans le chapitre 2, nous avons également présenté deux approches statistiques (profil circulaire et *vector scoring*) utilisées pour interpréter les mesures interpersonnelles. Le profil circulaire affiche les scores interpersonnels sur les huit octants du circumplex. Le *vector scoring* indique le comportement interpersonnel prédominant [Wiggins et al., 1988, Gurtman and Balakrishnan, 1998, Gurtman, 2009a, Locke and Adamic, 2012].

D'autre part, les travaux précédents ont souligné que les comportements humains sont naturellement multimodaux et séquentiels: nous interagissons les uns avec les autres par le biais de multiples canaux de communication (parole, regard, geste, etc.). De plus, ces comportements sont coordonnés dans le temps. Notre objectif est de comprendre comment ces comportements sont coordonnés aux moments critiques, les patterns séquentiels qu'ils présentent et leur association avec différentes attitudes interpersonnelles. Ainsi, dans notre travail, nous avons choisi de représenter les variations d'attitudes comme des séquences de signaux non-verbaux *multimodaux* et *temporels*.

Etat de l'art sur la modélisation de l'attitude sociale pour les agents virtuels

Le sujet général de cette thèse est de concevoir un ACA capable d'adapter son attitude envers l'utilisateur en fonction de son rôle et du contexte de l'interaction. Par exemple,

il devrait être dominant avec un candidat à une offre d'emploi et amical avec un enfant autiste. Dans le chapitre 3, nous présentons une vue d'ensemble des travaux les plus pertinents qui sont en rapport avec notre sujet: la modélisation d'attitude pour les agents virtuels. Nous nous concentrons également sur les travaux basés sur la séquentialité et la temporalité du comportement non-verbal en tant que aspect important de la modélisation du comportement de l'humain ou de l'agent. Les travaux existants qui modélisent les attitudes des ACAs traitent différentes questions : quels comportements de l'ACA influencent le plus la perception de son attitude? Comment l'attitude d'un ACA change-t-elle dans le temps? Comment générer automatiquement le comportement d'un ACA en fonction de son attitude interpersonnelle? A partir de l'aperçu des travaux existants, nous avons conclu que la combinaison du comportement verbal et non-verbal conduit à une meilleure reconnaissance des attitudes [Bee et al., , Callejas et al., 2014, Chollet et al., 2017]. Cependant, seuls quelques travaux ont fait ce choix [Chollet et al., 2014b, Cafaro et al., 2016b]. La plupart des travaux se basent sur le circumplex interpersonnel pour représenter les attitudes et considèrent les deux dimensions d'attitude à la fois. Généralement ces modèles reposent sur l'expression non-verbale des attitudes. Cependant, aucun travail n'exploite les informations temporelles de ces comportements. Notre travail surmonte cette limitation en prenant en compte les informations temporelles (moment de déclenchement et durée) des comportements non-verbaux.

Certains chercheurs ont souligné l'importance de la séquentialité et la temporalité des comportements non-verbaux pour mieux modéliser les attitudes. Dans le chapitre 3, nous présentons les travaux existants qui englobent la séquentialité du comportement non-verbal afin de comprendre et de prédire des phénomènes tels que l'émotion et l'attitude interpersonnelle. La plupart des travaux existants considèrent uniquement l'ordre des signaux tout en ignorant leurs informations temporelles [Chollet et al., 2014b, With and Kaiser, 2011]). Certaines travaux considèrent un nombre limité de modalités [Fricker et al., 2011, With and Kaiser, 2011, Yu et al., 2010, Zhang and Boyles, 2013]. Seuls quelques uns de ces travaux ont exploré les séquences extraites de comportements humains pour générer les comportements des agents virtuels [Chollet et al., 2014a]. Notre travail traite toutes ces limitations en prenant en compte les informations temporelles de comportements humains. Nous proposons un modèle entièrement automatique, séquentiel, temporel et génératif pour extraire et générer des séquences non-verbales qui représentent des variations d'attitude.

Sequence Mining: état de l'art et notre algorithme

Dans notre travail, nous représentons une variation d'attitude sous forme de séquences de signaux non-verbaux. En nous basant sur un algorithme de sequence mining, nous extrayons d'un corpus multimodal les séquences de comportements les plus pertinentes caractérisant une variation d'attitude. Sequence mining est une tâche d'exploration de données qui vise à découvrir les patterns pertinents cachés dans une grande base de séquences. Un pattern est une sous-séquence qui se produit fréquemment dans l'ensemble de données. Sequence mining a connu à un large éventail d'applications réelles dans de nombreux do-

maines, tels que l'analyse des tickets de caisse, NLP, la bioinformatique et l'analyse du comportement humain [Chollet et al., 2017, Fricker et al., 2011]. Par exemple, dans le contexte de l'analyse des tickets de caisse, l'exploration de séquence peut être utilisée pour identifier les séquences d'articles fréquemment achetés par les clients. Cela peut être utile pour comprendre le comportement d'achat des clients lors de la prise de décisions marketing. GSP est l'algorithme d'exploration de séquence le plus répandu [Srikant and Agrawal, 1996]. En prenant en entrée une base de séquences et un seuil de fréquence minimum (f_{min}), GSP découvre des patterns fréquents en se basant sur l'ordre des signaux. Par exemple, à partir du jeu de données $\{ABB, ABC, CABA, CABCA\}$ avec f_{min}=2, les patterns fréquents de taille trois sont $\{CAA, ABA, ABC, CAB\}$. Cependant, le fait de considérer que l'ordre des événements peut devenir une limitation lorsque des informations temporelles sont importantes, tels que: quel est le délai entre deux événements temporels? A quel moment se déclenche un événement temporel? Et quelle est sa durée? Pour pallier ce problème, des algorithmes de sequence mining temporel sont conçus pour traiter les informations liés au temps. Dans notre travail, nous nous concentrons sur ces algorithmes car ils permettent de répondre à nos questions de recherche: étant donné le contexte actuel (défini par les signaux non-verbaux précédents), (i) à quel moment un signal non-verbal doit-il se déclencher? Et (ii) quelle est sa durée?

Les algorithmes existants de sequence mining temporal ont été évalués en utilisant des données synthétiques. Par conséquence, ils ne parviennent généralement pas à gérer efficacement les données réels. L'algorithme que nous proposons, HCApriori (Hierarchical Clustering Apriori), surmonte les principales limitations des algorithmes existants: il est entièrement automatique et il augmente l'homogénéité des clusters en implémentant une technique de clustering adaptée (hiérarchique). La motivation derrière le développement de HCApriori est résumée dans les points suivants:

- **Meilleure gestion de la parcimonie (sparsity) des données**: la parcimonie est un aspect important à prendre en compte lors du clustering de données. Les travaux existants ne se sont pas concentrés sur ce phénomène car ils ont été appliqués à des données synthétiques (artificielles), souvent non exposées à ce problème.

- **Meilleure homogénéité des clusters**: lors de l'utilisation des algorithmes de partitionnement pour le clustering (comme Kmeans), les événements regroupés peuvent être très éloignés dans le temps. Pour surmonter cette limitation, notre algorithme HCApriori s'appuie sur une méthode de classification hiérarchique qui impose un minimum de similarité aux événements du même groupe;

- **Entièrement automatique**: les algorithmes existants nécessitent que l'utilisateur fournisse une estimation a priori du nombre de clusters ou de leur emplacement approximatif. Grâce à la classification hiérarchique, HCApriori ne nécessite aucune saisie manuelle. Toutes les étapes de l'algorithme sont entièrement automatiques;

- **Seuil de dissimilarité personnalisé et calculé automatiquement**: dans les travaux précédents, le seuil de dissimilarité ϵ est défini par l'utilisateur et il a la même

valeur pour tous les types d'événements. Cependant, cette contrainte peut être restrictive car la durée des événements peut différer considérablement selon le types d'événement. Par exemple, la durée d'un sourire sera probablement beaucoup plus courte que la durée d'une posture. HCApriori offre la possibilité de personnaliser ce paramètre pour chaque type d'événement. De plus, il peut être difficile de définir manuellement la valeur ϵ pour chaque type d'événement en raison du nombre de types d'événements (27 dans notre corpus) présents dans les données et également en raison des différences entre les types d'événements. Pour résoudre ce problème, nous proposons un moyen efficace de configurer ϵ automatiquement.

HCApriori fonctionne en deux étapes: (1) une classification hiérarchique est d'abord appliquée pour fusionner des signaux dans le même cluster si et seulement si leur distance temporelle est inférieure à ϵ. À la fin de cette étape, le centroïde de chaque cluster représente un pattern de longueur un. (2) En prenant en entrée les résultats de l'étape précédente, une procédure semblable à l'algorithme Apriori [Rakesh Agrawal, 1994] est adaptée pour générer des patterns temporels plus longs. En comparant HCApriori à l'état de l'art, nous avons constaté que notre algorithme surmonte de manière significative les algorithmes existants en termes de précision d'extraction. HCApriori est open source et il disponible sur github[1]. Nous avons amélioré les métriques standard, *support* et *confidence*, qui reflètent la qualité des patterns extraits. Notez que ces métriques sont à l'origine uniquement basées sur la fréquence d'occurrence des événements. Nous les avons étendues en intégrant le critère de temporalité pour une évaluation plus pertinente.

Modélisation séquentielle de la variation d'attitude

L'objectif de notre travail est de développer un agent virtuel capable d'exprimer des variations d'attitude en fonction du contexte de l'interaction. Par exemple, il devrait pouvoir augmenter son niveau de dominance lorsqu'il interroge un candidat à une offre d'emploi. Les attitudes interpersonnelles sont exprimées par des comportements non-verbaux (par exemple, regard, expression faciales, mouvements de tête, etc.). En outre, les attitudes ne sont pas seulement exprimées par des signaux spécifiques, mais aussi par la dynamique de ceux-ci (ordre et temporalité). Par conséquence, nous représentons une variation d'attitude comme une séquence temporelle de signaux non-verbaux dans laquelle chaque signal est défini par un moment de déclanchement et une durée. Pour la modélisation de la variation des attitudes, nous utilisons un corpus d'entretien d'embauche où un recruteur peut exprimer attitudes différentes envers le candidat. Ce corpus a été annoté à deux niveaux: comportement non-verbal et attitude interpersonnelle des recruteurs [Chollet et al., 2014b]. Plusieurs modalités de comportement non-verbal ont été annotées telles que le regard, les mouvements de tête, les expressions faciales, etc. L'annotation de la dominance et de l'amicalité a été réalisée de manière continue. Chaque annotateur annote une seule dimension d'attitude à la fois (dominance ou amicalité) et la valeur de

[1] https://github.com/dermosamo/HCApriori.git

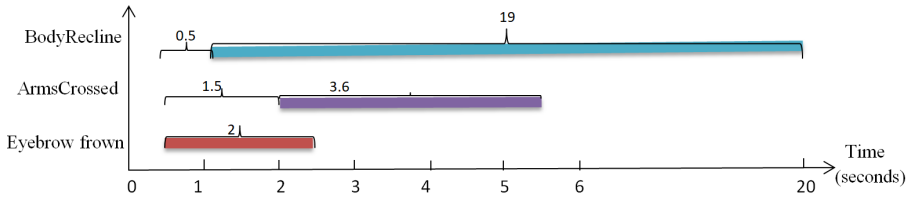


Figure 2 – Exemple d'un pattern représentant une augmentation de dominance.

l'annotation varie entre -1 et 1. En utilisant ce corpus, nous avons segmenté le comportement non-verbal du recruteur en se basant sur les variations d'attitudes. Pour chaque variation qui se produit lorsque les recruteurs parlent, nous récupérons tous les signaux non-verbaux apparaissant durant cette variation. Ces signaux composent une séquence de signaux non-verbaux. Par exemple, une séquence temporelle représentant une augmentation de dominance est composée d'un hochement vertical de la tête de 4 à 6 secondes suivi par un croisement des bras de 5 à 9 secondes. Cette segmentation nous permet de construire quatre bases de séquences de comportements non-verbaux représentant quatre types de variation d'attitude: augmentation et diminution de dominance ainsi que augmentation et diminution d'amicalité. Nous extrayons également des séquences de comportements qui apparaissent durant une attitude "neutre". Nous définissons par une expression "neutre" d'attitude les segments du corpus annotés avec une valeur d'attitude autour de zéro. Nous appelons ces séquences extraites "référence". Ainsi, nous obtenons des séquences représentant respectivement une "dominance neutre" et une "amicalité neutre". L'étape de segmentation donne des bases de séquences temporelles associées aux variations attitudes et aux attitudes "neutres". L'objectif est d'extraire les patterns les plus pertinents pour chaque variation d'attitude. Nous effectuons cette extraction en utilisant notre algorithme HCApriori. Un patterns représentant une augmentation de dominance est illustré à la figure 2, il peut être interprété comme suit: 1,5 s. avant que le recruteur augmente sa dominance, il fronce ses sourcils pendant 2 secondes. Pendant ce temps, il croise ses bras pendant 4.6 secondes tout en se penchant en arrière.

Afin d'évaluer les patterns non-verbaux extraits avec notre modèle, nous conduisons une expérience perceptive. Nous évaluons quatre différentes catégories de patterns non-verbaux dénotant quatre variations d'attitude: augmentation de dominance (*DomInc*), diminution de dominance (*DomDec*), augmentation d'amicalité (*FrInc*) et diminution d'amicalité (*FrDec*). Pour chaque condition, nous évaluons quatre patterns non-verbaux. Nous évaluons également un pattern exprimant une dominance "neutre" et un pattern exprimant une amicalité "neutre". À l'aide de la plate-forme d'agent virtuel appelée GRETA-VIB [Pecune et al., 2014], nous générons des vidéos montrant un agent simulant certains patterns, sélectionnés de manière aléatoire à partir du l'ensemble des patterns extraits. Nous produisons ainsi un total de 18 vidéos: 16 **vidéos de comparaison** (4 variations d'attitude $\times$ 4 patterns) et deux **vidéos de référence**: dominance "'neutre" (notée *DomRef*) et amicalité "neutre" (notée *FrRef*). Nos hypothèses sont:

- **H.Ref**: pour *DomRef* et pour *FrRef*, l'ACA sera évaluée comme exprimant une attitude "neutre";
- **H.Dom**: pour *DomInc*, l'ACA sera évalué comme **plus dominant** par rapport à l'ACA dans *DomRef*
- **H.Sub**: pour *DomDec*, l'ACA sera évalué comme **plus soumis** par rapport à l'ACA dans *DomRef*;
- **H.Fr**: pour *FrInc*, l'ACA sera évaluée comme **plus amical** par rapport à la valeur l'ACA dans *FrRef*;
- **H.Hos**: pour *FrDec*, l'ACA sera perçue comme **plus hostile** par rapport à l'ACA dans *FrRef*.

Résultats et discussion

Nous analysons les résultats de trois manières différentes: 1) en traçant les résultats sur le circumplex interpersonnel, 2) en réalisant des tests statistiques et 3) en calculant le taux de reconnaissance des variations d'attitude. Les vidéos de référence sont générées à partir des séquences non-verbales annotées avec des valeurs d'attitude proches de zéro. Nous avons supposé que l'agent dans ces vidéos serait perçu comme exprimant une attitude "neutre". À notre grande surprise, le résultat de l'étude montre que l'agent est évalué comme amical ce qui rejette l'hypothèse **H.Ref**. Nous ne trouvons aucune différence significative dans la perception de l'agent dans les vidéos de référence et dans la condition *FrInc*. Par conséquence, l'hypothèse **H.Fr** n'est pas validée. Une explication pourrait être que, dans la mesure où l'agent dans les vidéos de référence est déjà évalué comme amical, l'agent dans le *FrInc* n'est pas perçu comme étant nettement plus amical que dans les vidéos de comparaison (*FrRef*). Les trois autres hypothèses, **H.Dom**, **H.Sub** et **H.Hos**, sont validées.

Selon la représentation des attitudes sur le circumplex interpersonnel, les deux pôles d'une dimension d'attitude (dominance vs. soumission et amicalité vs. hostilité) sont symétriques par rapport au centre du circumplex. En conséquence, on s'attendait à ce que l'augmentation de l'attitude vers un pôle donné se traduise par une diminution de la perception du pôle opposé. Par exemple, une augmentation de l'amicalité diminuerait la perception de l'hostilité et inversement. Sur la base du profil circulaire, cette relation est observée pour les deux pôles de chaque dimension d'attitude, dans les deux sens des variations d'attitude.

Plusieurs travaux sur la modélisation d'attitude reposent sur l'hypothèse qu'il existe un effet de compensation entre les deux dimensions d'attitude. Pour calculer quelles attitudes sociales un agent transmet à son interlocuteur, des traveaux ont défini des règles telles que les émotions positives ressenties par l'agent, augmentent son amicalité et diminuent sa dominance envers l'utilisateur. Inversement, les émotions négatives diminuent son amicalité et augmentent sa dominance [Kasap et al., 2009, Pecune et al., 2016]. D'autres travaux s'appuient sur la théorie de complémentarité interpersonnelle dans l'interaction [Ravenet

et al., 2015]. Selon cette théorie, deux personnes devraient exprimer des attitudes complémentaires ou anti-complémentaires afin de maintenir une interaction: exprimer des attitudes similaires sur la dimension d'amicalité et des attitudes opposées sur la dimension de dominance [Leary, 1957, Kiesler, 1996]. Mais, à notre connaissance, il n'existe aucune étude, en termes de perception, sur l'interrelation des dimensions des attitudes interpersonnelles. Pour étudier cette interrelation, nous avons évalué les deux dimensions des attitudes en même temps. Cela permet de souligner un effet de compensation entre la perception de dominance et d'amicalité tirée des observations suivantes:

- L'augmentation de dominance conduit à une perception de diminution d'amicalité;
- La diminution de dominance conduit à une perception d'augmentation de l'amicalité;
- La diminution d'amicalité conduit à une perception d'augmentation de la dominance.

Nous observons qu'il existe une forte corrélation entre l'augmentation de la dominance (*DomInc*) et la diminution d'amicalité (*FrDec*). Une explication est que certains signaux non-verbaux ont le même effet sur la perception de domination et d'hostilité [Knutson, 1996, Tiedens et al., 2000, Carney et al., 2005, Ravenet et al., 2013]. Par exemple, la dominance et l'hostilité sont toutes les deux caractérisées par une expression faciale négative [Knutson, 1996, Tiedens et al., 2000, Carney et al., 2005, Ravenet et al., 2013]. Trois de nos 5 hypothèses (**H.Dom**, **H.Sub** et **H.Hos**) ont été validées. Les séquences exprimant les variations d'attitude correspondantes sont donc correctement reconnues. Cela confirme notre hypothèse selon laquelle les variations d'attitude peuvent être représentées par des séquences de signaux non-verbaux ordonnés dans le temps. Notre prochaine étape consiste à utiliser les séquences extraites pour planifier une variation d'attitude d'un agent virtuel.

Planification des attitudes pour des agents virtuels

La plate-forme GRETA-VIB a été développée pour soutenir la création d'ACA socio-émotionnelles [Pecune et al., 2014]. Sur cette plate-forme, l'agent affiche des énoncés avec des fonctions de communication et des états émotionnels. L'attitude de l'ACA joue un rôle essentiel pour la réalisation de l'objectif d'interaction [Kasap et al., 2009, Ochs et al., 2010, Pecune et al., 2016]. Dans le Chapitre 6, nous décrivons comment nous avons amélioré la plate-forme GRETA-VIB avec notre modèle d'attitude. À cette fin, nous avons développé un planificateur d'attitudes qui combine la variation d'attitude de l'ACA avec ses intentions communicatives. GRETA-VIB est basée sur le framework SAIBA [Vilhjalmsson et al., 2007] dont l'architecture est illustrée sur la figure 3. Tout d'abord, le *intent planner* génère les intentions communicatives de l'ACA (ce que l'agent a l'intention de communiquer). Les intentions communicatives sont représentées dans le langage FML (Functional Markup Language) [Heylen et al., 2008]. Ensuite, le *behavior planner* traduit ces intentions communicatives en un ensemble de signaux multimodaux (par exemple gestes, expressions faciales). Enfin, ces signaux sont transformées en animation finale de

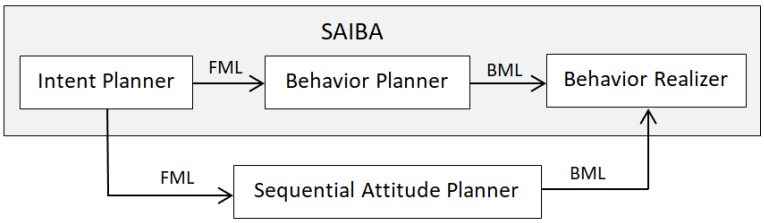


Figure 3 – L' architecture SAIBA avec le nouveau module: *Sequential Attitude Planner*.

l'ACA par le *behavior realizer*. Pour modéliser l'attitude sociale, nous remplaçons le *behavior planner* par un nouveau appelé *Sequential Attitude Planner*. Il prend en entrée un fichier FML (contenant l'énoncé à dire par l'agent), les intentions et la variation d'attitude que l'ACA exprimera envers l'utilisateur. Le modèle de planification d'attitude est composé de quatre étapes:

1. **Génération d'une séquence d'intention:** le planificateur d'intentions communicatives génère une séquence de comportements non-verbaux exprimant les intentions communicatives contenues dans le fichier FML d'entrée. Par exemple, l'intention communicative *saluer* peut être exprimée soit par un *geste* soit par une *expression faciale (sourire ou haussement des sourcils)*. Une fois que toutes les intentions communicatives ont été instanciées, nous obtenons une séquence de comportements multimodaux que nous appelons *séquence d'intention* (S_{int}).

2. **Sélection d'une séquence d'attitude:** une fois les intentions communicatives instanciées, l'étape suivante consiste à choisir la séquence la plus appropriée traduisant la variation d'attitude souhaitée. La pertinence est définie ici comme la séquence la plus représentative pour exprimer la variation d'attitude et comme la plus similaire, en termes de présence et temporalité de comportements multimodaux, à la *séquence d'intention* (S_{int}). Pour cela, parmi les patterns extraits avec notre algorithme HCApriori, nous sélectionnons une *séquence d'attitude* (S_{att}) qui convient le mieux à ces deux propriétés. Dans cette étape, nous associons également les comportements de S_{int} aux comportements de S_{att}.

3. **Enrichissement de la séquence d'intention:** afin de calculer la séquence finale de comportements que l'agent affichera pour communiquer ses intentions avec une variation d'attitude, nous enrichissons la séquence S_{int} avec l'ensemble de comportements d'attitude. Cela correspond à la fusion des séquences d'intention et d'attitude: chaque comportement dans la séquence S_{att} qui n'apparaît pas dans la séquence d'intention est ajouté à S_{int}.

4. **Remplacement des signaux:** afin de représenter la relation entre les comportements non-verbaux et les variations d'attitude, nous calculons la fréquence d'occurrence d'un comportement donné b par rapport à une variation d'attitude donnée V. Nous considérons qu'un comportement b_1 est plus représentatif d'une variation d'attitude V qu'un comportement b_2 si la fréquence d'occurrence de b_1 est supérieure à la

fréquence d'occurrence de b_2. Enfin, notre modèle remplacera chaque comportement b_{int} de S_{int} par son comportement associé b_{att} dans S_{att} si la fréquence de b_{att} est supérieure à la fréquence de b_{int}.

Evaluation et résultats

Nous concevons une expérience empirique dans laquelle les participants comparent un ensemble de paires de vidéos. Chaque paire est composée d'une vidéo du recruteur virtuel sans variation d'attitude et d'une vidéo avec une variation d'attitude. Nous choisissons sept questions qui ont un contenu verbal plutôt «neutre». Voici un exemple de question: *si nous décidons de vous proposer ce travail, quand seriez-vous prêt pour commencer?*. Sept vidéos de référence (*ref*) sont générées sans notre planificateur d'attitude (c'est-à-dire il n'y a aucune variation d'attitude) et 28 avec notre planificateur d'attitude (4 variations d'attitude $\times$ 7 questions). Pour l'évaluation, nous suivons la même procédure que l'évaluation précédente. Nous évaluons cinq conditions expérimentales qui relatives aux quatre variations d'attitude: augmentation de dominance (*DomInc*), diminution de dominance (*DomDec*), augmentation d'amicalité (*FrInc*), diminution d'amicalité (*FrDec*) ainsi que l'attitude de référence (*Ref*). Nous avons les mêmes hypothèses que la première expérience.

Contrairement à notre première étude, notre hypothèse **H.DomDec** n'a pas été validée. Chollet et ses collègues [Chollet et al., 2014b] ont obtenu des résultats similaires pour leur recruteur virtuel simulant une diminution de dominance. Ce résultat peut être lié au contexte de l'interaction où l'agent joue le rôle de recruteur. Dans un tel contexte, le recruteur a tendance à contrôler l'interaction et apparaît donc naturellement dominant et pas soumis. En outre, l'agent dans *DomDec* est perçu comme plus hostile et moins amical que l'agent de la vidéo de référence (*DomRef*), alors que dans la première étude, il est perçu comme plus amical et moins hostile. Ce changement de perception confirme l'importance du contexte d'interaction susceptible de modifier la perception d'une attitude. Un autre résultat qui peut être lié au rôle de l'agent est la corrélation positive entre les dimensions d'amicalité et de dominance: une augmentation d'amicalité conduit à une augmentation de dominance. Cependant, dans la première étude, l'augmentation de l'amicalité était corrélée avec la diminution de dominance.

Une dimension d'attitude est représentée par deux pôles symétriques (dominance/-soumission, amicalité/hostilité). Nous nous attendions à une relation négative entre les deux pôles d'une dimension d'attitude (une augmentation d'un pôle donné entraînerait une diminution de la perception du pôle opposé). Cette hypothèse est statistiquement significative pour l'amicalité/ l'hostilité: lorsque l'agent est perçu comme plus hostile, il est également perçu comme étant moins amical, et inversement. Pour l'autre dimension, il existe une forte tendance: lorsque l'agent est évalué comme plus dominant, il est également perçu comme étant moins soumis, et inversement. Nous avons conclu de cette étude que notre modèle de planification d'attitudes permet à l'ACA d'exprimer une variation d'attitudes, en particulier une augmentation de dominance et une diminution d'amicalité. La diminution de dominance n'a pas été reconnue. Ce résultat pourrait être provoqué par le contexte d'interaction, ici le rôle de l'agent. La perception de

l'amicalité de l'agent dans les conditions de référence semble affecter la reconnaissance de l'augmentation d'amicalité.

Modèle génératif des comportements de l'agent dans l'interaction humain-agent

Dans l'interaction humain-humain, les humains adaptent leur comportement en fonction du comportement de leurs interlocuteurs [Burgoon et al., 2010]. Par exemple, un interlocuteur hoche la tête pour indiquer qu'il est d'accord avec l'orateur, il regarde le même objet ou sourit en réponse au sourire de son locuteur. Dans ce contexte, notre objectif est de modéliser un agent capable d'adapter son comportement en fonction du comportement de l'utilisateur. Les comportements non-verbaux jouent un rôle important dans le maintien de l'engagement entre l'utilisateur et l'agent dans les interactions humain-agent [Fong et al., 2002, Arai and Hasegawa, 2004, Breazeal., 2004, Woolf and Burleson, 2009]. C'est pourquoi nous sommes particulièrement intéressés par une adaptation dynamique des comportements non-verbaux de l'agent à ceux de son interlocuteur. Pour adapter le comportement de l'agent en fonction de celui de l'utilisateur, nous tirons parti des avancées récentes dans le domaine des réseaux de neurones, en particulier un type de réseau très répandu appelé LSTM. Cette approche englobe simultanément la séquentialité et la temporalité du comportement non-verbal au fil du temps. Le modèle conçu adopte une approche réactive pour prévoir en permanence le comportement de l'agent en réponse au comportement de l'utilisateur. Il prend en entrée à la fois le comportement passé de l'utilisateur et de l'agent et prédit le comportement prochain de l'agent. Plus précisément, il prédit le sourire, les mouvements de tête et le regard de l'agent. Pour intégrer et évaluer notre modèle LSTM appelé IL-LSTM (Interaction Loop LSTM), nous créons un système d'interaction dans lequel l'agent interagit en temps réel avec un utilisateur humain. Le système prend en entrée les données de l'utilisateur, calcule ce que l'agent doit dire ainsi que l'animation correspondante. À notre connaissance, notre modèle est la première tentative de produire, en temps réel, le sourire, les mouvements de tête et la direction du regard pour un agent virtuel en considérant le sourire, les mouvements de tête et la direction du regard à la fois de l'agent et de l'utilisateur, ainsi que les intentions de communication de l'agent. Notre système d'interaction se décompose en 4 modules illustrés sur Figure 4 décrits dans les sections suivants.

1. **EyesWeb: Analyse du comportement de l'utilisateur**
   EyesWeb XMI est une plate-forme open source permettant d'enregistrer et d'analyser le comportement humain en temps réel [Volpe et al., 2016]. En utilisant EyesWeb, nous extrayons les expressions faciales, les mouvements de tête, le regard et l'activité vocale de l'utilisateur.

2. **Flipper: Gestion des dialogues**
   Au cours d'une interaction humain-agent, l'agent doit choisir le prochain acte de dialogue en fonction de l'évolution de l'interaction avec l'utilisateur. Par exemple,

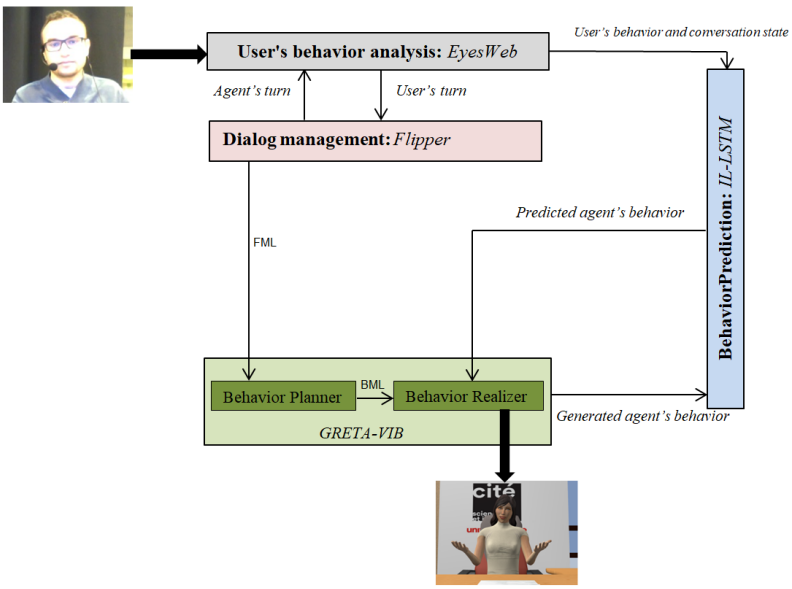


Figure 4 – L'architecture de notre système.

l'agent commence l'interaction en saluant l'utilisateur. Une fois que l'utilisateur a répondu, l'agent se présente et demande le nom de l'utilisateur, etc. Pour sélectionner les actes de dialogue les plus appropriés de l'agent, nous utilisons Flipper [van Waterschoot et al., 2018]. Flipper permet de spécifier des règles de dialogue pour les agents virtuels. Dans notre système, Flipper gère le tour de parole de l'agent et envoie des messages à EyesWeb pour indiquer le moment où l'agent commence et finit de parler. Inversement, EyesWeb envoie à Flipper l'activité vocale de l'utilisateur. En fonction de l'étape de dialogue, Flipper sélectionne le discours suivant de l'agent.

3. **GRETA-VIB**
   GRETA-VIB est utilisée comme plate-forme d'agent virtuel [Pecune et al., 2014]. Dans notre système, Flipper génère les intentions communicatives de l'agent et envoie le comportement multimodal correspondant au format FML au *behavior planner*.

4. **BehaviorPrediction: Prédiction du comportement de l'agent**
   Nous souhaitons adapter le comportement de l'agent en réponse au comportement de l'utilisateur. En d'autres termes, l'agent doit non seulement communiquer ses intentions, mais également adapter son comportement en temps réel par rapport au comportement de l'utilisateur. Pour atteindre cet objectif, nous devons prédire le comportement de l'agent à chaque instant. Ceci est réalisé en ajoutant un module spécifique à l'architecture du système. En tant que tel, le module "BehaviorPrediction" est chargé de calculer les comportements adaptatifs.

A chaque pas de temps t (c'est-à-dire pour chaque image), le module *BehaviorPrediction* calcule les paramètres d'animation faciale et corporelle de l'agent pour le pas de temps $t + 1$ (c'est-à-dire, l'image suivante). Ce module prend en entrée les paramètres d'animation faciale et corporelle de l'utilisateur et de l'agent, ainsi que l'état de la conversation sur une fenêtre temporelle de 20 image (*frame*). La prédiction de la prochaine image est faite par le modèle IL-LSTM. Ainsi, l'animation prédite de l'agent est calculée à partir des animations précédentes de l'utilisateur et de l'agent.

Les comportements de l'utilisateur sont extraits à l'aide de EyesWeb. EyesWeb communique avec le modèle IL-LSTM en envoyant les comportements de l'utilisateur et l'état de la conversation. Flipper envoie le fichier FML au *behavior planner*. Ce dernier calcule le comportement multimodal de l'agent et l'envoie au *behavior realizer* qui calcule l'animation de l'agent. Ensuite, avant d'envoyer chaque image au générateur d'animation, nous fusionnons l'animation calculée à partir des intentions communicatives avec l'animation prédite pour adapter le comportement de l'agent à celui de l'utilisateur. Nous répétons cette opération à chaque image. Ceci est fait au niveau du *behavior realizer* avec l'entrée du module *BehaviorPrediction*.

Evaluation

Afin d'évaluer notre modèle génératif, nous conduisons une expérience interactive dans laquelle notre module *BehaviorPrediction* est utilisé pour générer automatiquement le comportement de l'agent en tenant compte du comportement de l'utilisateur. L'agent, nommé Alice, joue le rôle d'un guide virtuel décrivant une exposition sur les jeux vidéo destinée aux visiteurs de la cité des sciences et de l'industrie à paris. Nous supposons que l'adaptation le comportement d'Alice à celui de l'utilisateur augmentera l'engagement et la satisfaction de celui-ci. Notre objectif est de vérifier si notre modèle qui adapte le comportement de l'agent améliore l'interaction et l'expérience de l'utilisateur. Pour évaluer l'engagement de l'utilisateur, Van Vugt *et al.* ont proposé un questionnaire pour évaluer le comportement de l'agent en fonction des dimensions du model PEFiC [van Vugt et al., 2006]. À partir de ce questionnaire, nous adaptons un ensemble d'adjectifs permettant de mesurer la perception de l'agent en termes de réalisme, de compétence et de pertinence. On mesure également l'engagement et la satisfaction de l'utilisateur. De plus, afin de mesurer l'amicalité perçue de l'agent, nous avons utilisé quatre éléments du questionnaire IAS [Wiggins, 1979]. Enfin, pour mesurer l'attitude a priori des participants à propos des agents virtuals, nous adoptons huit questions du questionnaire NARS [Nomura et al., 2006]. 101 participants ont pris part à notre expérience, 50% sont des femmes et 95% sont français. Les résultats ont montré que les utilisateurs étaient en effet plus satisfaits de leur interaction avec Alice lorsqu'elle adaptait son comportement. Cependant, ces résultats n'étaient significatifs que lorsqu'Alice adaptait son sourire. Un biais lié à l'utilisateur aurait pu empêcher d'avoir des résultats significatifs pour les autres expressions (regard et mouvement de tete).

Prediction d'engagement

Dans le Chapitre 8, nous nous sommes concentrés sur un aspect important de l'interaction humain-agent: l'engagement. L'engagement assure que l'interaction se déroule sans perte d'intérêt ni de motivation de la part de l'utilisateur. Après avoir étudié les comportements qui contribuent le plus à un changement dans la perception de l'engagement, nous nous sommes concentrés sur les expressions faciales, les mouvements de tête et la direction du regard. Ces trois comportements représentent des indicateurs pertinents d'engagement. Nous avons développé un modèle basé sur LSTM pour prédire le niveau d'engagement de l'utilisateur. Le model a été entrainé à partir de la base de données NoXi contenant des conversations entre expert et novice. Nous avons exploré la contribution de différentes caractéristiques (*features*) multimodales, à savoir le regard, les mouvements de tête et les unités d'action (*action units*), à la prédiction de l'engagement. Les résultats ont révélés que les unités d'action contribuaient davantage que les mouvements de tête et le regard à la prédiction de l'engagement. Nous avons également étudié l'importance de prendre en compte le comportement de l'interlocuteur pour prédire l'engagement de locuteur. Les résultats ont souligné l'importance de prendre en compte le comportement des deux partenaires dans une interaction dyadique pour la prédiction de l'engagement. Notre modèle a été intégré dans une plate-forme ECA, ce qui permet de prédire, en temps réel, le niveau d'engagement de l'utilisateur.

Contents

1 Introduction **1**

1.1 Context and Research Issues 1

1.2 Contributions 3

1.3 Manuscript Organization 4

1.4 Publications 5

2 Theoretical Background **7**

2.1 Attitude Definition 7

2.2 Attitude Representation 8

2.3 Interpersonal Circumplex Measurements and Interpretation 9

2.3.1 Measurements 11

2.3.2 Scoring and Interpreting the IPC Measurements 13

2.4 Multimodal Expressions of Social Attitude 15

2.5 Non-verbal Behavior Interpretation 17

3 State of the Art on Attitude Modeling for Virtual Agents **21**

3.1 Attitude Modeling for Embodied Conversational Agents 21

3.1.1 ECA's Behavior Expressing Attitude 22

3.1.2 Attitude Dynamics over Time 24

3.1.3 Attitude Generation Models 25

3.2 Sequence-Based Multimodal Behavior Modeling 26

4 Sequence Mining: State of the Art and Our Algorithm **31**

4.1 Non-Temporal Algorithms 32

4.2 Temporal Algorithms 33

4.3 Temporal Sequence Mining Algorithms 35

4.4 HCApriori Algorithm 37

4.4.1 Definitions 38

4.4.2 Algorithm 39

4.4.3 Evaluation and Results 41

4.5 Pattern Quality Assessment 43

4.6 Conclusion 45

5 Sequence-Based Attitude Variation Modeling **47**
5.1 Extraction of Relevant Patterns Expressing Attitude Variations 48
5.1.1 Building Sequence Databases Representing Attitude Variations 48
5.1.2 Pattern Extraction 54
5.2 Evaluation of the Extracted Patterns 54
5.2.1 Experimental Design 55
5.2.2 Measures 55
5.2.3 Hypotheses 56
5.2.4 Results 57
5.2.5 Discussion 63
5.3 Conclusion 65

6 Attitude Planner **67**
6.1 GRETA-VIB 68
6.2 Sequential Attitude Planner Model 70
6.2.1 Intention Sequence Generation 70
6.2.2 Attitude Sequence Selection 71
6.2.3 Intention Sequence Enrichment 72
6.2.4 Signal Replacement 73
6.3 Evaluation 73
6.3.1 Experimental Design 74
6.3.2 Results 74
6.4 Discussion 77
6.5 Conclusion 79

7 Generative Model of Agent's Behaviors in Human-Agent Interaction **81**
7.1 Related works 83
7.2 Corpus 84
7.3 Neural Networks and LSTM 87
7.3.1 Neural Networks: Overview 87
7.3.2 Recurrent Neural Networks: LSTM 88
7.4 LSTM Model 89
7.4.1 Prediction Model 89
7.4.2 Evaluation 91
7.5 Architecture 92
7.5.1 EyesWeb: User's Behavior Analysis 93
7.5.2 Flipper: Dialog Management 93
7.5.3 GRETA-VIB 95
7.5.4 BehaviorPrediction: Agent's Behavior Prediction 96
7.6 Evaluation 97
7.6.1 Independent Variables 97

7.6.2 Measures . . . 97
7.6.3 Hypotheses . . . 98
7.6.4 Protocol . . . 100
7.7 Results . . . 100
7.8 Conclusion . . . 102

8 Engagement Modeling in Dyadic Interactions **105**
8.1 Related Works . . . 106
8.1.1 Engagement-Related Behaviors . . . 106
8.1.2 Engagement Prediction . . . 106
8.2 LSTM Model for Engagement Prediction . . . 108
8.2.1 Data . . . 108
8.2.2 Model . . . 109
8.2.3 Results . . . 112

9 Conclusion **117**
9.1 Summary . . . 117
9.1.1 Attitude Variation Modeling . . . 117
9.1.2 Adapting Agent's Behavior According to the User's Behavior . . . 119
9.1.3 Engagement Prediction . . . 120
9.2 Limits and Perspectives . . . 120

Appendices **141**

A Results of the First Study **143**

B Results of the Second Study **147**

C Engagement Prediction **151**

List of Tables

2.1 IPC Measures . . . 13
2.2 Representative adjectives of each IPC octant for several inventories. . . . 13
2.3 Non-verbal signals involved in the expression of interpersonal attitudes. . . 17

3.1 A comparison between attitude models. . . . 26
3.2 A comparison between sequence-based behavior modeling methods. . . . 29

4.1 Transaction database. . . . 33
4.2 Mean duration of some non-verbal signals in seconds. . . . 37
4.3 A comparison of temporal sequence mining algorithms. . . . 37
4.4 Example value of ϵ customized by event's type (non-verbal signals). . . . 42

5.1 Annotated non-verbal behaviors in Tardis corpus. . . . 50
5.2 Mean and standard deviations of attitude variation attributes. . . . 50
5.3 Size of sequences for each attitude variation occurring when the recruiters are speaking. . . . 53
5.4 Size of extracted patterns for each attitude variation as well as the two "reference" attitudes. . . . 54
5.5 Example of patterns obtained with HCApriori. . . . 54
5.6 Selected adjectives from ICL and IAS. . . . 56
5.7 Angle and vector length of the agent for all conditions. . . . 59
5.8 The four evaluated patterns for *DomInc*. . . . 60
5.9 Friedman test and Bonferroni post-hoc test exploring the effects of the different patterns on *DomInc*. . . . 62
5.10 Recall, precision, and F-measure for each attitude variations. . . . 62
5.11 Distribution of the predictions over the actual conditions. The predictions highlighting the compensation effects given in Section 5.2.4.3 and the friendliness perception of the agent in *DomRef* and *FrRef*. . . . 63
5.12 Relationships between attitude variations and attitude perception. . . . 64

6.1 Angle and vector length of each condition. . . . 76
6.2 Recall, precision, and F-measure for the multimodal model. . . . 77
6.3 Distribution of the predictions over the actual conditions. . . . 78
6.4 Relationships between attitude variations and attitude perception. . . . 78

7.1 An overview of works related to adaptation model. . . . 84

7.2 Percentage of each conversation state in NoXi database. 86
7.3 Performance of our model in comparison to the baseline. 92
7.4 Items used to evaluate the perception of the agent. 99
7.5 Items used for measuring user's engagement and satisfaction as well as user's apriori attitude toward virtual agent. 99
7.6 Distribution of participants w.r.t their age. 101
7.7 Cronbach's α, mean and standard deviation of each measured dimension for the five conditions. 101

8.1 An overview of works related to engagement prediction. 107
8.2 Percentage of each engagement level in NoXi database for expert and novice. 109
8.3 Prediction of expert engagement based on different models. 113
8.4 Prediction of expert's engagement level using each feature separately (in addition to their union) for the three different configurations. 114

A.1 Mean, standard deviation of variables, and distribution of participants' answers for *DomRef*. 143
A.2 Mean, standard deviation of variables, and distribution of participants' answers for *FrRef*. 144
A.3 Mean, standard deviation and frequency of participants' answers for *DomInc*. 144
A.4 Mean, standard deviation and frequency of participants' answers for *DomDec* in the first experiment. 145
A.5 Mean, standard deviation and frequency of participants' answers of variables for *FrInc*. 145
A.6 Mean, standard deviation and frequency of participants' answers of variables for *FrDec* in the first experiment. 146

B.1 Mean, standard deviation and frequency of participants' answers for the seven reference videos. 147
B.2 Mean, standard deviation and frequency of participants' answers of variables for *DomInc*. 148
B.3 Mean, standard deviation and frequency of participants' answers of variables for *DomDec* 148
B.4 Mean, standard deviation and frequency of participants' answers of variables for *FrInc*. 149
B.5 Mean, standard deviation and frequency of participants' answers of variables for *FrDec* 149

C.1 Prediction of engagement level using several features and expert LSTM. . . 151
C.2 Prediction of engagement using several novice LSTM. 152
C.3 Prediction of engagement using dyadic LSTM. 152

List of Figures

1 The interpersonal circumplex (IPC). xiii
2 Exemple d'un pattern représentant une augmentation de dominance. xvii
3 L' architecture SAIBA avec le nouveau module: *Sequential Attitude Planner*. . xx
4 L'architecture de notre système. xxiii

1.1 An interaction featuring a set of non-verbal signals [Vinciarelli et al., 2009] 3

2.1 The interpersonal circumplex (IPC). 9
2.2 Examples of interpersonal behaviors plotted on the Interpersonal Circumplex [Isbister, 2006]. 10
2.3 Classification of interpersonal behavior in Leary's model [Leary, 1957]. . . . 12
2.4 Circular profiles of two persons P_1 and P_2 that participated to the CSIE inventory [Locke, 2012]. The profile of P_1 is represented by a dashed line. . 14
2.5 Representation of a prototypical embarrassment response. The mean duration of each action is equal to the beginning of the interval with the leftmost edge of the photography and ending with the end of the arrow [Keltner, 1995]. 18

3.1 Examples of an ECA showing non-verbal behavior (control: no behavior, dominant: akimbo (hands on the hip), submissive: neck-adaptor, cooperative: head tilt right and non-cooperative: gaze aversion) [Straßmann et al., 2016]. 22
3.2 Beattie's taxonomy of interruption types (image from [Cafaro et al., 2016b]). 23
3.3 Screenshot of the online graphical application designed to annotate the behavior of the ECA for a given attitude [Ravenet et al., 2013]. 25

4.1 Temporal data model. 32
4.2 Allen's interval relationships [Kam and Fu, 2000]. 34
4.3 Different representations of a temporal pattern [Kam and Fu, 2000]. 34
4.4 Process of temporal sequence mining algorithms. 35
4.5 Synthetic data vs. real data representing gaze signals. Real data is very sparse while synthetic data can easily be grouped into three clusters. 36
4.6 Accuracies of the compared algorithms for different values of f_{min}. Here, ϵ is used for evaluation and its value is less than the one used for pattern extraction. 42

4.7 Accuracies of the compared algorithms for different values of f_{min}. Here, ϵ is used for evaluation and its value is less than the one used for pattern extraction. . . . 43
4.8 Accuracies of the compared algorithms for different values of f_{min}. Here, ϵ is used for evaluation and its value is less than the one used for pattern extraction. . . . 44
4.9 Running time evolution with respect to the dataset size. . . . 45
4.10 Representation of overlap between two temporal events. . . . 45

5.1 Annotation scale of interpersonal attitude (dominance dimension) [Chollet, 2015]. . . . 49
5.2 Attitude variation illustration. . . . 49
5.3 Non-verbal behavior segmentation based on attitude variations. . . . 49
5.4 Boxplots of starting value (v^{vs}) for each attitude variation. . . . 51
5.5 Boxplots of ending value (v^{ve}) for each attitude variation. . . . 51
5.6 Boxplots of v^{value} for each attitude variation. . . . 52
5.7 Boxplots of $v^{duration}$ for each attitude variation. . . . 52
5.8 Distribution of non-verbal signals w.r.t. attitude variation. . . . 53
5.9 Pattern representing dominance increase. . . . 54
5.10 A screenshot from CrowdFlower evaluation platform. . . . 57
5.11 Interpersonal adjectives from ICL and IAS used in our experiment and their placement in the interpersonal circumplex. . . . 57
5.12 Plotting the ipsatized scores for *DomInc*, *DomDec* and *Ref*. . . . 58
5.13 Circular profile of the agent for *FrInc*, *FrDec* and *Ref*. . . . 59
5.14 Differences in the ECA perception between the comparison video and the reference video. The blue color indicates that the variable is evaluated as being more expressed in the reference video; the green indicates that the variable is evaluated as being more expressed in the comparison video; and the black denotes no difference between the reference video and the comparison video. . . . 60
5.15 Boxplots of some variables of the four patterns expressing dominance increase . (1) “strongly disagree”, (2) “partially disagree”, (3) “neutral”, (4) “partially agree”, (5) “strongly agree”. . . . 61

6.1 SAIBA architecture enhanced with the new integrated module: *Sequential Attitude Planner*. . . . 68
6.2 Example of a FML file. . . . 68
6.3 “Behavior set” for question mark *boundary-HH*. . . . 69
6.4 Example of a BML file. . . . 69
6.5 GRETA-VIB platform. . . . 70
6.6 Outline of the sequential attitude planning model. . . . 71
6.7 An illustrative example for the sequential attitude planning model. . . . 72
6.8 Time adjustment of b_{att} when $b^e_{att} > b^s_{int}$ . . . 73
6.9 Time adjustment of b_{att} when $b^e_{att} < b^s_{int}$ . . . 73

6.10 Plotting octant scores on the IPC for dominance increase and decrease. . . . 75
6.11 Octant scores for friendliness increase and decrease. 76
6.12 Differences in the ECA perception between the comparison video and the reference video. The blue color indicates that the variable is evaluated as being more expressed in the reference video; the green indicates that the variable is evaluated as being more expressed in the comparison video; and the black denotes no difference between the reference video and the comparison video. 77

7.1 The most frequently-discussed topics in NoXi corpus. 85
7.2 Expert-novice interaction. 85
7.3 Example of annotations of the conversation states detected from Expert and Novice's voice activity. 86
7.4 Action Units corresponding to the activation of different facial muscles. . . . 86
7.5 A simple neural neural network architecture. 88
7.6 LSTM cells. 89
7.7 Our neural network graph with multiple outputs. 91
7.8 Plots of ground truth and predicted agent's smile and head rotations (x and y) through time. 92
7.9 Architecture of our system. 93
7.10 EyesWeb interface. Relying on OpenFace, EyesWeb extracts the user's facial expressions. 94
7.11 Information state of our system. 94
7.12 Example of Flipper template. 95
7.13 SAIBA architecture. 95
7.14 PEFiC model [van Vugt et al., 2006] 98
7.15 Our system for human-agent interaction 100

8.1 Three LSTM models for predicting expert's engagement level. 110
8.2 Training and validation loss (top) and accuracy (bottom) w.r.t. the number of epochs. 111
8.3 F-measure w.r.t the sequence length given as input. 112

9.1 An example of discrete annotation of attitude using the IPC with three levels for each octant. 122

Chapter 1

Introduction

Contents

1.1 Context and Research Issues 1
1.2 Contributions 3
1.3 Manuscript Organization 4
1.4 Publications 5

HUMANS use different modalities while interacting with each other, including speech, gesture, facial expression, body postures, etc. These modalities provide cues about emotions, personality, and intention among other functions. Designed on the human model, Embodied Conversational Agents (ECAs) are virtual characters that can interact autonomously with human beings by imitating their natural behaviors. Toshiba's *Yoko* and Airbus's *Tim* are examples of animated virtual agents that answer technical and commercial questions of customers. In the recent years, much research effort has focused on the improvement of human-agent interaction experience, especially in the last years where virtual agents became more and more present in our everyday lives. They can be used for a variety of applications ranging from education and training to therapy [Nojavanasghari et al., 2016, Chollet et al., 2017, Nojavanasghari and Hughes, 2017]. With the growing interest in human-agent interactions, it is desirable to make these interactions pleasant and human-like. In the context of this thesis, we aim at enhancing the interaction experience between humans and ECAs.

1.1 Context and Research Issues

The different components of human behavior have been extensively studied by psychologists, sociologists, and more generally cognitive scientists, since many decades [Leary,

1957, Argyle, 1988, Burgoon et al., 2010]. More recently, the works from cognitive science brought this domain to the reach of artificial intelligence, with, for example, the advent of Embodied Conversational Agents (ECAs)[1]. A large body of research focused on endowing ECAs with human-like abilities and behaviors, like expressing emotions, social attitudes, and reacting properly to the user's actions, for example, to keep her engaged during the interaction. In the context of this thesis, we aim at improving the interaction between human users and virtual agents. To this end, we develop computational models for endowing ECA with the capacity to: (1) express different social attitudes in accordance with the interaction context, (2) adapt its behavior according to the user's behavior and its communicative intentions, (3) predict the user's engagement level during human-agent interaction. Our goal is to enrich the state-of-the-art with more adapted and fine-grained models and algorithms.

Humans continuously express different social attitudes toward each other depending on the interaction context that includes factors such as the interaction partner, role, personality, goal, etc. For example, a person may show a kind of *dominance* in some work contexts while being warm when going out with friends. The same person will not behave in the same way in these different circumstances. She will not display the same behaviors. She may use a more formal language at work, show a more upright posture, smile less, while she may laugh and gesture expressively with friends and family. In this context, we aim at endowing a virtual agent with the capacity to express different social attitude depending on the interaction context. For example, the ECA should be friendly with a customer when answering her question but more dominant with a job candidate to train her preparing the job interview.

For a deeper understanding of attitude expression, we first need to explore what makes a person appear more/less dominant or more/less friendly? That is, finding out what patterns of non-verbal behaviors trigger a change (variation) in the perception of social attitudes. Such an analysis should build on the associations and the dynamics (temporal variation) of non-verbal behavior which is well informative for characterizing and interpreting attitudes (see Figure 1.1).

On the other hand, humans tend to adapt their behavior throughout the interaction in accordance with the behavior of the interaction partners [Burgoon et al., 2010]. For example, a listener nods to indicate agreement with the speaker, she gazes the same object or smiles in response to the interlocutor's smile. In light of this, virtual agents should adopt a user-in-the-loop approach and change their behavior in response to the user's actions and behaviors. Such a dynamic interaction would benefit to maintain user's engagement in the interaction. In the recent years, engagement modeling has gained increasing attention due the important role it plays in human-agent interaction. The agent should be able to detect, in real time, the engagement level of the user in order to react accordingly. In this context, our goal is to develop a computational model to predict engagement level of the user in real time. Relying on previous findings, we use facial expressions as predictive

[1]The term Embodied Conversational Agent is used, instead of Virtual Agent, when it is capable of engaging in conversation with one another or with humans. In this document, we use both terms interchangeably.

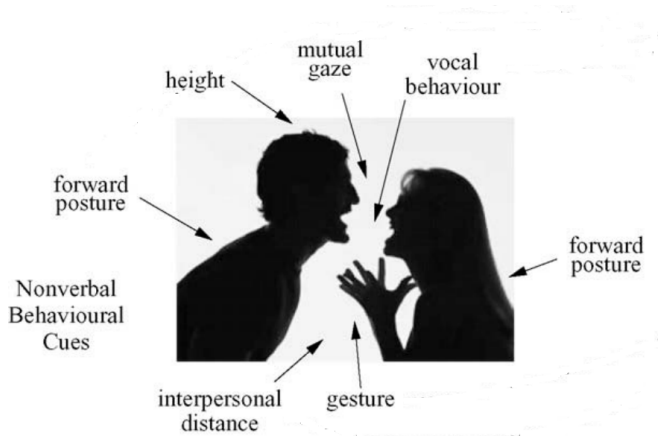


Figure 1.1 – An interaction featuring a set of non-verbal signals [Vinciarelli et al., 2009]

features [Allwood and Cerrato, 2003, Castellano et al., 2009c]. Moreover, engagement is not only measured from single cues, but from the combination of several cues that arise over a certain time window [Peters et al., 2005, Bickmore et al., 2012]. Thus, for better engagement prediction, we should consider the variation of facial expressions over time.

1.2 Contributions

In this thesis, we are interested in generating the non-verbal behavior of virtual agents including the attitude variation it should express, as well as the behavior adaptation in response to the user's behavior. We also focus on user's engagement level prediction during human-agent interaction. To achieve these goals, we encompass the dynamics of non-verbal behavior relying on appropriate techniques: *temporal sequence mining* and *recurrent neural networks (LSTM)*. Specifically, our main contributions can be summarized in the following three points:

Attitude variation generation for virtual agents: the novelty of our proposition is to model attitude variations as sequences of non-verbal signals, while considering the temporal characteristics (starting time and duration) of these signals. Thus, we develop a fully-automatic, sequential, temporal, and generative model for extracting and generating non-verbal sequences representing attitude variations. Our practical contributions to the domain of attitude modeling are:

- Designing a new temporal sequence mining algorithm. As the existing algorithms of temporal sequence mining have been intrinsically designed for synthetic data, they generally fail to efficiently deal with real-word data. We propose an algorithm, called HCApriori, that addresses the main limitations of existing algorithms for modeling real data. The conducted experiments show a significant improvement of HCApriori over the state-of-the-art algorithms.

- Modeling attitude variation: we consider signal's temporality for attitude modeling. We rely on HCApriori algorithm to extract, from multimodal corpus, relevant temporal sequences expressing attitude variations. A perceptive study is conducted to validate the expressivity of extracted sequences.

- Enriching the ECA platform (GRETA-VIB [Pecune et al., 2014]) with an attitude planner allowing virtual agents to simultaneously express attitude variations and other communicative intentions (e.g., performative, emphasis). Based on the extracted sequences (characterizing attitude) as well as the communicative intentions of the agent, we generate the final non-verbal behaviors that should be displayed by the agent. A perceptive experiment is conducted to evaluate the perception of an ECA communicating while displaying the attitudes generated with our attitude planner.

Real-time ECA's behavior adaptation: the human behavior is naturally dynamic. It is the result of summing up different factors: communicative intents, interaction context, responses to the other interaction partner's behavior... To adapt the agent's behavior according to the user's one, we take advantage of the recent advances in the domain of neural networks, specifically a popular type of networks called LSTM. This approach simultaneously encompasses the sequentiality and temporality of behavior. The designed model adopts a user-in-the-loop approach to constantly generate the behavior of the agent in response to the user's behavior. To our knowledge, this is the first attempt to produce real time facial expressions for virtual agents driven from both agent's and user's behaviors as well as agent's communicative intentions.

User's engagement prediction: In this part of the thesis, we shed light on an important aspect of human-agent interaction: *engagement*. Engagement ensures the interaction to go on without loss of interest or motivation. The agent should be able to continuously detect the engagement level of the user in order to react in a proper way. To this end, we develop a LSTM-based model to predict, in real time, the engagement level of the user.

1.3 Manuscript Organization

Chapter 2 lays the foundations for the works of this thesis. It gives a theoretical background on attitude definition, its expression, representation, and interpretation. In Chapter 3, we address previous efforts that are related to ours, by discussing attitude models for ECAs, and existing works that model the sequentiality of multimodal behaviors. Chapters 3—6 are all related to our first contribution, the modeling of social attitude as sequences of multimodal behaviors. Sequence mining task and our algorithm (HCApriori) are presented in Chapter 4. Chapter 5 describes our methodology for attitude variation modeling as a sequence of non-verbal signals. The sequences are integrated in an ECA platform to automatically generate the attitude variation of ECAs as presented in Chapter 6.

Our second contribution (behavior adaptation) is addressed in Chapter 7. After giving a quick overview on neural networks and LSTM, we describe the architecture of our behavior adaptation model along with the obtained results.

Our third contribution (engagement prediction) is presented in Chapter 8. After giving an overview on engagement-related behaviors and existing works for engagement prediction, we describe our model. Finally, in Chapter 9 we conclude this thesis and give some future perspectives.

1.4 Publications

- **Journal**
  Dermouche, S. and Pelachaud, C. (2019). Sequential Attitude Planner for Virtual Agents. IEEE Transactions on Affective Computing. **Under review**

- **International conferences**
  Dermouche, S. and Pelachaud, C. (2019). Engagement Modeling in Dyadic Interaction. In proceedings of the 21th ACM International Conference on Multimodal Interaction. Suzhou, Jiangsu, China.

  Dermouche, S. and Pelachaud, C. (2019). Generative Model of Agent's Behaviors in Human-Agent Interaction. In proceedings of the 21th ACM International Conference on Multimodal Interaction. Suzhou, Jiangsu, China.

  [Mancini et al., 2019] Mancini, M., Biancardi, B., Dermouche, S., Lerner, P., and Pelachaud, C. (2019). Managing Agent's Impression Based on User's Engagement Detection. In International Conference on Intelligent Virtual Agents (IVA'19). Paris, France.

  [Dermouche and Pelachaud, 2018a] Dermouche, S. and Pelachaud, C. (2018). Attitude Modeling for Virtual Character Based on Temporal Sequence Mining. In Proceedings of the 5th International Conference on Movement and Computing, pages 1–8, Genoa, Italy.

  [Cafaro et al., 2017] Cafaro, A., Wagner, J., Baur, T., Dermouche, S., Torres, M. T.,Pelachaud, C., André, E., and Valstar, M. (2017). The NoXi Database: Multimodal Recordings of Mediated Novice-Expert Interactions. In proceedings of the 19th ACM International Conference on Multimodal Interaction, pages 350–359, Glasgow, Scotland.

  [Dermouche and Pelachaud, 2018c] Dermouche, S. and Pelachaud, C. (2018c). From Analysis to Modeling of Engagement as Sequences of Multimodal Behaviors. In Proceedings of the 11th Language Resources and Evaluation Conference (LREC

2018), pages 786–791,Miyazaki, Japan.

[Dermouche and Pelachaud, 2018b] Dermouche, S. and Pelachaud, C. (2018c). Expert-Novice Interaction: Annotation and Analysis. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan.

[Dermouche and Pelachaud, 2016b] Dermouche, S. and Pelachaud, C. (2016). Sequence-based multimodal behavior modeling for social agents. In proceedings of the 18th ACM International Conference on Multimodal Interaction, pages 29–36, Tokyo, Japan.
[Dermouche, 2016] Dermouche, S. (2016). Computational Model for Interpersonal Attitude Expression. In proceedings of the 18th ACM International Conference on Multimodal Interaction, pages 554–558, Tokyo, Japan.

- **National conferences**
  [Dermouche et al., 2018] Dermouche, S. and Pelachaud, C. (2018). Extraction and Evaluation of Non-Verbal Signal Sequences Expressing Interpersonal Attitudes. WACAI 2018, Porquerolles, France

  [Dermouche and Pelachaud, 2016a] Dermouche, S. and Pelachaud, C. (2016). Extraction and Evaluation of Non-Verbal Signal Sequences Expressing Interpersonal Attitudes. WACAI 2016, Brest, France

Chapter 2

Theoretical Background

Contents

2.1 Attitude Definition 7
2.2 Attitude Representation 8
2.3 Interpersonal Circumplex Measurements and Interpretation 9
2.3.1 Measurements 11
2.3.2 Scoring and Interpreting the IPC Measurements 13
2.4 Multimodal Expressions of Social Attitude 15
2.5 Non-verbal Behavior Interpretation 17

ATTITUDE is an essential component of human-human interaction. Because it is complex and subject to different interpretations, attitude has attracted attention from various research fields. In this Chapter, we focus on the theoretical bases and definitions to clarify the concept of human attitude and help understating the underlying models. From a computational perspective, we describe the existing frameworks for attitude representation and measurement. These frameworks differ according to the context of study (e.g., interpersonal problems, personal traits, etc.). Finally, we investigate the relation between non-verbal behavior and the perceived attitude. We show how the key characteristics of non-verbal signals, such as sequentiality ans temporality, are determinant to correctly perceiving the conveyed attitudes.

2.1 Attitude Definition

Interpersonal, or social attitudes have captured much attention in social psychology [Scherer, 2005], social linguistics [Biber, 2006] and, more recently, in social signal processing [Ballin et al., 2004, Lee and Marsella, 2011, Ravenet et al., 2015, Chollet et al., 2017, Janssoone,

2016]. Consequently, different definitions of attitude, also called *stance*, have been proposed [Scherer, 2005, Du Bois, 2007, Chindamo et al., 2009]. From Linguistics we can cite the definition of Biber: *"personal feelings, attitudes, judgments, or assessments that a speaker or writer has about the information in a proposition"* [Biber, 2006]. Another definition is given by Du Bois: *"Stance is a public act by a social actor, achieved dialogically through covert communicative means, of simultaneously evaluating objects, positioning subjects (self and others), and aligning with other subjects, with respect to any salient dimension of the socio-cultural field"* [Du Bois, 2007]. Thus, expressing attitude consists of evaluating an object, positioning w.r.t a situation of a person and also aligning with other persons.

Attitudes are defined by Chindamo as: *"The expressive side of a stance includes unimodal as well as multimodal vocal or gestural (in a wide sense including all communicative and informative body movements) verbal or nonverbal contributions"* [Chindamo et al., 2009]. According to same source, attitudes are expressed through both verbal and -not less importantly - non-verbal behaviors. Moreover, Argyle reported that *"non-verbal signals have a much greater impact than equivalent verbal signals in communicating interpersonal attitude"* [Argyle, 1988] (page 85).

Klaus Scherer defined attitude as *"an affective style that spontaneously develops or is strategically employed in the interaction with a person or a group of persons, coloring the interpersonal exchange in that situation"* [Scherer, 2005]. The important aspect of attitudes underlined in this definition is that attitudes are *dynamic*: an attitude is an an affective style that *"colors"* an interaction. Then, attitudes are not only expressed by a given signal at a certain time but rather by the coordination and dynamics of a series of multimodal signals whose meaning arises from the interrelation of interactants' behaviors.

From all definitions, we can conclude that attitudes are *interpersonal*, *multimodal* and *dynamic*. In our work, we focus on interpersonal attitudes, i.e., attitudes that are expressed toward a person, in particular, the attitude that our virtual agent will express toward the user. The term *multimodal* means that attitudes are expressed verbally and non-verbally. According to Argyle, both modalities contribute equally to attitude expression. Hence, we are are interested by the non-verbal expression of attitude. Finally, attitudes change over time. Our goal is to encompass the dynamics of attitudes by jointly considering sequentiality and temporality of non-verbal signals.

2.2 Attitude Representation

Attitudes can be formally represented in different ways. Burgoon proposes a representation of social attitudes along twelve dimensions [Burgoon and Hale, 1984]. The first seven dimensions are independent: dominance (dominance or submission), arousal (degree of emotional arousal and responsiveness), relaxation (degree of composure and relaxation exhibited), similarity (resemblance between parties), formality, social orientation, and intimacy. The last five are related to the concept of intimacy: trust, superficiality of the relationship, hostility, involvement, and inclusion. Some other models represent attitudes in a small number of dimensions. For example, in [Schutz, 1958], the author considers the dimensions of inclusion, control and affection. Heise characterizes the social attitudes ac-

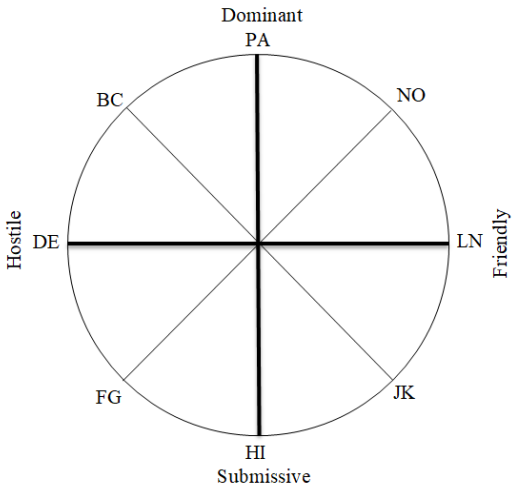


Figure 2.1 – The interpersonal circumplex (IPC).

cording to the dimensions of evaluation (positive vs negative), power (powerful vs weak) and activity (excited vs relaxed) [Heise, 2010]. In [Barbulescu et al., 2015], the authors rely on Mind Reading taxonomy to represent attitudes [Rajendran, 2004]. This taxonomy is composed of 412 emotions grouped into 24 categories such as: seductive, jealous, surprised, responsible, etc.

Attitudes can be represented from other perspectives. Argyle used a graphical representation called Interpersonal Circumplex (IPC), where attitudes are plotted according two orthogonal axes (see Figure 2.1): a vertical axis for dominance and a horizontal axis for friendliness. These two dimensions have initially been used in 1957 by Leary [Leary, 1957]: *"in surveying the list of more or less generic interpersonal trends, it became clear that they all had some reference to a power or affiliation factor. When dominance-submission was taken as the vertical axis and hostility-affection as the horizontal, all of the other generic interpersonal factors could be expressed as combinations of these four nodal points"* (page, 64) [Leary, 1957]. Thus, each interpersonal or social behavior, like "forceful", can be represented, within the IPC, as a weighted combination of dominance and friendliness. Figure 2.2, plots some examples of social behaviors on the IPC.

2.3 Interpersonal Circumplex Measurements and Interpretation

As a reminder, our goal is to develop a generative model of attitude variations for a virtual agent. We will evaluate our model by the perception of the generated attitudes through a questionnaire. To find the most relevant adjectives that characterize the perception of attitude, we perform a literature review on the usage of IPC measurements. This study will help us fill the evaluation requirements.

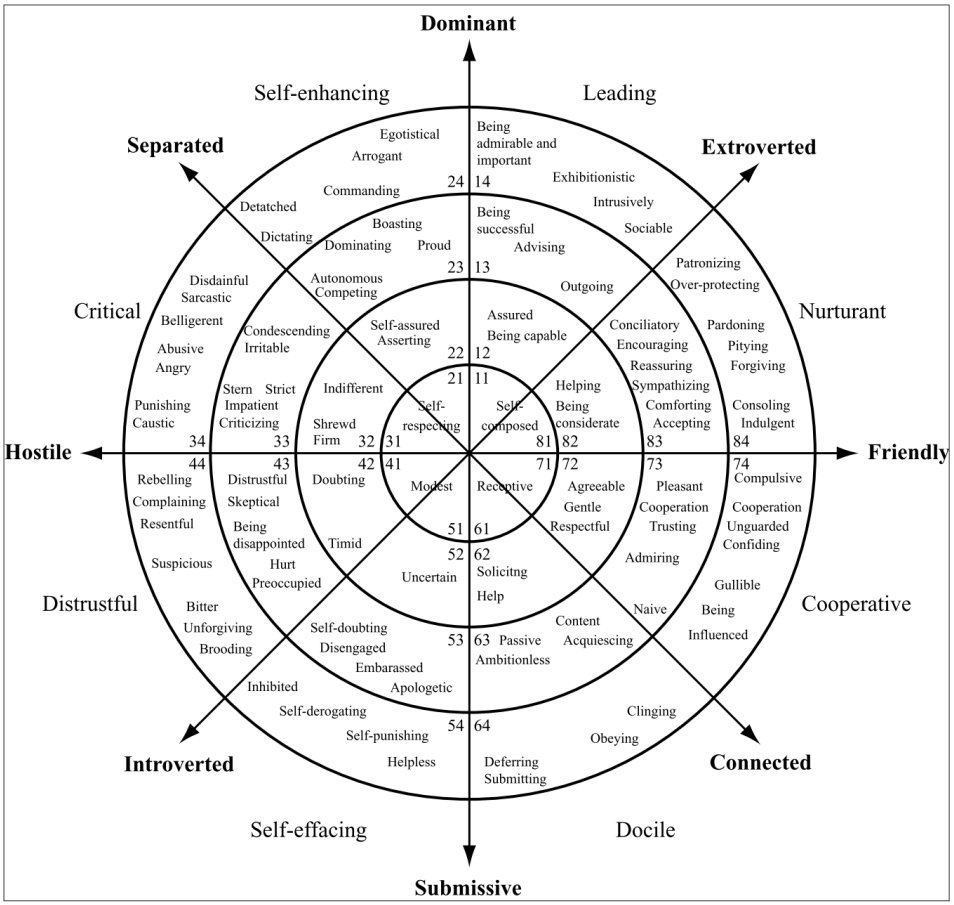


Figure 2.2 – Examples of interpersonal behaviors plotted on the Interpersonal Circumplex [Isbister, 2006].

2.3.1 Measurements

The IPC has recently become a popular model for assessing interpersonal dispositions such as interpersonal problem (e.g., problems related to assaulting others) [Alden et al., 1990], value (how interpersonal experiences, such as expressing herself openly, are important to a person?) [Locke, 2000], self-efficacy (interpersonal actions a person believes she can express) and traits (e.g., firm) [Wiggins, 1995]. All the IPC measures are based on the same theory: there is a particular location within the circumplex space for each interpersonal disposition. Most IPC inventories split the IPC into eight octants or scales that are alphabetically labeled counterclockwise: PA, BC, DE, FG, HI, JK, LM and NO (see Figure 2.1). Each octant can be represented by a set of characteristic adjectives, e.g., *dominant* and *assertive* for PA octant.

To build an IPC inventory, psychologists started by building the questionnaires describing the measured interpersonal dispositions, for example, by analyzing psychotherapy interviews. Then, using statistical analyses such as Principal Component Analysis, participant answers were clustered and displayed on the IPC. The works of Locke, Adamic, Acton, and Revelle provided overviews of interpersonal circumplex measures or inventories [Locke and Adamic, 2012, Acton and Revelle, 2014]. As they reported, the Interpersonal Check List (ICL), proposed by Leary, was the first ICP inventory [Leary, 1957]. Based on Sullivan's interpersonal theory of personality [Sullivan, 1953] on one hand, and observing interactions among psychotherapy group members on the other hand, Leary constructed a circumplex model that represented interpersonal traits (cf. Figure 2.3). As we can see, Leary classified 16 interpersonal behaviors on the interpersonal circle. Each of the 16 behaviors is evaluated by multi-level measures: (1) reflexes are illustrated in internal circle and indicated by alphabetical letters (*A* to *P*). (2) The center ring indicates the behaviors provoked by persons adapting the interpersonal behaviors. For example, a person who uses the reflex *P* tends to provoke others to *respect*. (3) The next circle illustrates extreme reflexes like *compulsive* and *dominant*. Finally, the circle perimeter is divided into eight interpersonal behaviors (e.g., managerial-autocratic). The ICL model has been widely used in psychological and socio-psychological research [Clark, T. L., & Taulbee, 1981]. However, researchers reported that ICL did not adequately fit the circumplex model. It presents significant measurement gaps between the four quadrants of the circumplex [Kiesler, 1996, Wiggins et al., 1988, Locke, 2000]. Wiggins *et al.* [Wiggins, 1979] proposed the Interpersonal Adjective Scales (IAS) to address these limitations.

An interpersonal adjective is defined as "*a pattern of dyadic interactions that has relatively clearcut social (status) and emotional (love) consequences for both participants (self and other)*" [Wiggins, 1979] (p. 398). Based on this definition, 800 terms were identified as interpersonal. For simplifying the rating and the interpretation, the 800 terms have been reduced to 128 adjectives and then to 64 in the final version of the IAS (IAS-R). To validate the IAS model, participants rated how accurately each adjective describes them on a 8-point scale. The methodology used to build IAS served as a basis for the development of other IPC measures like the Inventory of Interpersonal Problems (IIP). Moreover, IAS is now the standard measure of interpersonal traits [Locke and Adamic, 2012].

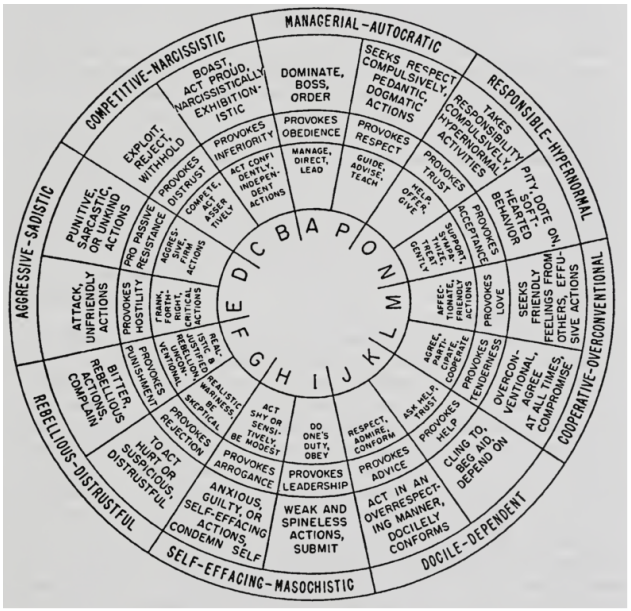


Figure 2.3 – Classification of interpersonal behavior in Leary's model [Leary, 1957].

Horowitz *et al.* studied a large sample of psychotherapy interviews for reporting the most frequent interpersonal problems [Horowitz, 1997]. Based on this study, the Inventory of Interpersonal Problems (IIP) was developed. It consists of 64 items that assess interpersonal excesses and deficiencies. participants rated how distressed they have been for each problem on a 5-point scale. The items are divided into two sections: "It is hard for me..." and "I am too much...". IIP was used to identify the relationship between interpersonal problems and psychopathology and psychotherapy [Ruiz et al., 2010]. Locke reported that the IIP can help guide therapeutic interventions for interpersonal problems [Locke and Adamic, 2012]. For example, the interpersonal problems assessed by the IIP are related to the types of interpersonal expectations that are readily targeted by therapeutic interventions. For example, dominant people expect others to be critical whereas friendly people expect others to be dismissive.

Self-efficacy is how confident a person is able to perform some action [Rogelberg, 2017]. The Circumplex Scales of Interpersonal Efficacy (CSIE) assess a person's confidence that she can successfully perform behaviors [Locke and Sadler, 2007]. Answers range from 0 (not at all confident) to 10 (absolutely confident).

Table 2.1 gives a summary of some IPC measures and table 2.2 indicates some representative adjectives for each octant of the IPC. For each measure, we indicate: the number of items used to measure a specific interpersonal disposition, the question and its answer scale, as well as the number and kind of participants who answered the questionnaire in the initial study.

Several other examples of ICP inventories exist, like the Interpersonal circumplex measures of interpersonal constructs [Gurtman, 2009b], the Support Actions Scale-Circumplex (SAS-C) [Trobst, 2000]), the Octant Scale Impact Message Inventory (SIMI) [Schmidt, J. A., Wagner, C. C., & Kiesler, 1999] and the Circumplex Scales of Interpersonal Values

Inventory	Year	Dispositions	Items	Question	Scale	Participants
ICL	1957	Reflexes	128	Respond if an item apply to you	5-point (adaptive to extreme)	
IAS	1979	Traits	64	Rate how each describe you	8-point (extremely inaccurate to extremely accurate)	1161 university students (British Columbia)
IIP	1997	Problems	64	Rate how distressing each problem has been	5-point (not at all to extremely)	200 patients
CSIV	2000	Value	64	When I am with him/her, it is important that...	5-point (not important to extremely important)	
CSIE	2007	Self-efficacy	32	Rate how confident you are	10-point (not at all confident to absolutely confident)	

Table 2.1 – IPC Measures

Measure	IMI-C	IAS-R	IIP-C	SAS	CSIE
LM	Friendly	Warm-Agreeable	Overly Nurturant	Nurturant	Dominant
NO	Friendly-Dominant	Gregarious-Extraverted	Intrusive	Engaging	Dominant-Distant
PA	Dominant	Assured-Dominant	Domineering	Directive	Distant
BC	Hostile-Dominant	arrogant-Calculating	Vindictive	arrogant	Yielding-Distant
DE	Cold-hearted	Cold	Critical	Yielding	
FG	Hostile-Submissive	Aloof-Introverted	Socially Avoidant	Distancing	Yielding-Friendly
HI	Submissive	Unassured-Submissive	Nonassertive	Avoidant	Friendly
JK	Friendly-Submissive	Unassuming-Ingenuous	Exploitable	Deferentia	Dominant-Friendly

Table 2.2 – Representative adjectives of each IPC octant for several inventories.

(CSIV) [Locke, 2000]. Most of existing IPC measures were developed for adults. More recently, specific inventories for children and adolescents have been developed, such as the Child and Adolescent Interpersonal Survey (CAIS) [Sodano and Tracey, 2007] and the Interpersonal Goals Inventory for Children (IGIC) [Ojanen et al., 2005]. Other IPC measures were developed for specific cultures, like the South African Personality Inventory [Hill et al., 2013] and the Dutch adjectives scales [Op Den Akker et al., 2013]. Although most IPC measures are used as self-report measures, they can be and have been used, with some changes to the instructions or items, to rate the behavior of specific targets (e.g., the virtual agent's behavior in a human-agent interaction [Locke, 2000]). For evaluating the attitude perception, previous works used IAS [Pecune, 2016, Cafaro et al., 2016a, Cafaro et al., 2016b, Janssoone, 2016] or ICL [Op Den Akker et al., 2013]. Inspired by these works, we rely on both IPC and IAS for evaluating the perception of an ECA expressing attitude variations (see. Chapter 5).

2.3.2 Scoring and Interpreting the IPC Measurements

After choosing a suitable IPC inventory depending at the task at hand, the next step is to score and interpret the answers of participants. One commune approach for analyzing such data is the circular profile. This profile presents each person's scores on the eight octants of the circumplex (cf. Figure 2.4). For computing this profile from any IPC inventory, Locke follows three steps [Locke, 2012]:

1. Compute the general factor score by averaging the eight octant scores;
2. Ipsatize octant scores by subtracting the general factor score from each octant score;

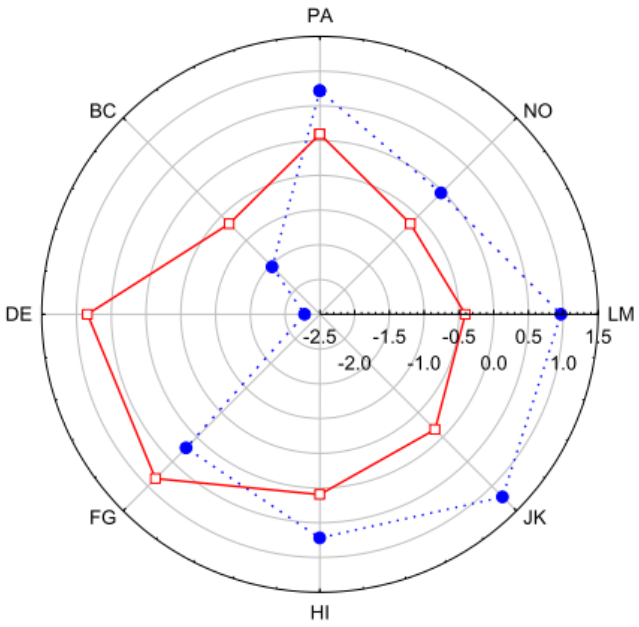


Figure 2.4 – Circular profiles of two persons P_1 and P_2 that participated to the CSIE inventory [Locke, 2012]. The profile of P_1 is represented by a dashed line.

3. Plot the ipsatized scores on the IPC ranging from the lowest value to the highest value.

To interpret the circular profiles, Gurtman explains: *"circular profiles tend to rise to a peak value and then decline. The peak clearly indicates the predominant trend in the profile and suggests the individual's predominant interpersonal style or typology"* [Gurtman, 2009b]. Figure 2.4 plots the circular profiles of two persons P_1 and P_2 who answered the CSIE inventory. For the profile P_1, the peak is in the lower-right region which suggests a friendly-yielding behavior, whereas for P_2 the peak is in the lower-left quadrant suggesting a hostile-yielding style. Based on the circular profile, we can also compare the behaviors of both participants P_1 and P_2: they are similar in efficacy for being dominant (PA) and yielding (HI). On the opposite, participant $P1$ is more friendly (LM) than distant (DE), unlike participant $P2$.

Leary introduced another approach called *vector scoring* by summarizing the circular profiles with a single point on the circumplex [Leary, 1957]: the vertical coordinate gives the perceived dominance based on Equation 2.1 whereas the horizontal coordinate characterizes the friendliness based on Equation 2.2, by combining the ipsatized octant scores as indicated in [Wiggins, 1979]. For example, the vertical coordinate (DOM) represents the weight of the octant scores according to their directions compared to the dimension of dominance. Thus, we sum the scores of PA, BC and NO (vary in the same direction as dominance) and we subtract the scores HI, FG and JK (vary in the opposite direction of dominance). The values of DOM and FR define a vector in the IPC space whose angle can be calculated by Equation 2.3 and length by Equation 2.5. The angle is adjusted as

indicated in Equation 2.4. The vector angle indicates the predominant interpersonal behavior [Wiggins et al., 1988, Gurtman and Balakrishnan, 1998, Gurtman, 2009a, Locke and Adamic, 2012].

$$DOM = 0.03\ (PA - HI) + 0.02\ (NO + BC - FG - JK) \tag{2.1}$$

$$FR = 0.03\ (LM - DE) + 0.02\ (NO - BC - FG + JK) \tag{2.2}$$

$$Angle = \tan^{-1} \frac{DOM}{FR} \tag{2.3}$$

$$AjustedAngle = \begin{cases} Angle + 0, & \text{if } DOM < 0 \text{ and } FR > 0 \\ Angle + 180, & \text{if } FR < 0 \\ Angle + 360, & \text{otherwise} \end{cases} \tag{2.4}$$

$$\text{Vector length} = \sqrt{DOM^2 + FR^2} \tag{2.5}$$

For example, the vector angle of person P_2 used above is 216° (in the FG octant) whereas the vector angle of person P_1 is in the JK octant (337°). Gurtman reports that *"a high vector length indicates a well-defined profile, with a clear central tendency; but low vector length suggests less definition to the profile and hence less confidence in any summary conclusion about the overall thematic trend in the personality"* [Gurtman, 2009a]. The vector length of P_1 in Figure 2.4 is many times greater than P_2's, which suggests that P_1 has a clearer interpersonal profile.

2.4 Multimodal Expressions of Social Attitude

Nonverbal behavior is an important component in human interaction. It can participate to the regulation of interaction (e.g., nodding may indicate an agreement with the speaker), it can complete and structure the speech (e.g., raising eyebrows can accentuate an element of the speech) [Ekman and Friesen, 1969, Argyle, 1988, Cosnier, 1997]. Non-verbal behavior also contributes to the expression of emotions and attitudes [Argyle, 1988]. In our work, we are interested in the expression of interpersonal attitude through non-verbal behavior. The relationship between non-verbal behaviors and attitudes has been widely investigated in psychology and sociology [Gifford, 1991, Mehrabian, 1969]. In the following, we summarize the most significant findings on the relation between non-verbal behaviors and interpersonal attitudes.

- **Gestures**: McNeill categorized gestures into two main categories: communicative gestures and adaptor gestures [McNeill, 1992]. Gestures can bear information about the speaker's attitude. For example, adaptor gestures that consist in touching oneself or manipulating objects are mainly related to submissive attitude but can also be associated with hostility in some cases [Burgoon, J. K. and Le Poire, 1999]. Touching

her interlocutor can be a sign of friendliness and of dominance depending on the type of the touch [Carney et al., 2005, Burgoon et al., 1984]. Frequency and expressivity of gestures, like amplitude and intensity, directly influence the perception of an attitude. Also, performing large gestures may be a sign of dominance. Dominant people are also generally characterized by gesturing more compared to submissive people.

- **Postures**: when two interacting persons adapt unconsciously their postures one to another, we can predict a change in their interpersonal attitudes [Richmond, V. & McCroskey, 2000]. Lafrance noted that postural mirroring (adopting the same posture as one's interlocutor) can be a sign of friendliness [Lafrance, 1982]. Leaning towards and taking a closer position to her interlocutor can be perceived as a sign of submission, whereas reverse behaviors, such as leaning backwards, could express dominance [Carney et al., 2005, Burgoon et al., 1984, Burgoon, J. K. and Le Poire, 1999]. Adopting a posture by occupying a large space, in the same way as large gestures, are signs of dominance [Carney et al., 2005, Burgoon et al., 1984, Burgoon, J. K. and Le Poire, 1999, Gifford and Hine, 1994]. For example, dominant people extend their legs more than submissive people do [Gifford, 1991].

- **Head direction and movement**: communicative functions of head movements are also varied. When listening, head movement can be a backchannel indicating an agreement, disagreement or understanding [Heylen et al., 2008]. Head direction and movement can also be relevant signals in predicting attitude. A bowed head can be a sign of submission, a head tilt of friendliness whereas a raised head may express dominance [Gifford, 1991, Debras and Cienki, 2012, Stivers, 2008]. On the other hand, a head shake can correlate to different attitudes: dominance [Gifford and Hine, 1994, Carney et al., 2005, Hall et al., 2005] and friendliness [Burgoon, J. K. and Le Poire, 1999, Gifford, 1991], depending on the context.

- **Gaze**: gaze is a crucial element in measuring social dimensions such as engagement [Kendon, 1967, Abele, 1986] and attitude [Argyle, M., Dean, 1965, Duncan, St. jr., Fiske, 1977, Burgoon et al., 1984, Hall et al., 2005]. Mutual gaze is a sign of dominance and friendliness whereas gaze shift is perceived as a sign of submission, while direct gaze is a sign of dominance. Generally, dominant people gaze more at their interlocutors than submissive ones [Hall et al., 2005].

- **Facial expression**: the role of facial expression and their impact on the perception of attitude have also been studied [Knutson, 1996, Tiedens et al., 2000, Carney et al., 2005]: joyful expressions are associated with friendliness and dominance, fearful and sadness expressions with submission, while anger and disgust expressions are linked to hostility and dominance. Smile has been reported as the typical signal of friendliness [Keating and al, 1981] but could also express dominance in some situations [Hall et al., 2005]. Finally, Keating *et al.* studied the influence of eyebrow movements on attitude perception and showed that, generally, a frown

eyebrow is perceived as expressing dominance while a raised eyebrow expresses submission [Keating and al, 1981].

Attitude	Associated non-verbal behaviors
Dominance	large gesture, touching others, leaning towards, raised head, head nod, head shake, eyebrow frown, smile, joy expression, mutual gaze
Submission	self-touch, manipulation of objects, leaning forwards, bowed head, gaze shift, fear and sadness expression, raised eyebrow
Friendliness	head tilt, head nod, smile, joy expressions, mutual gaze
Hostility	manipulation of objects, disgust expressions

Table 2.3 – Non-verbal signals involved in the expression of interpersonal attitudes.

In this Section, we focused on the correlation between non-verbal behavior and attitudes. Table 2.3 synthesizes some examples of non-verbal signals characterizing attitudes. In the next section, we present the different temporal aspects of non-verbal signals (such as starting time and duration) that could influence their perception, hence they must be considered when modeling the non-verbal behavior of virtual agents.

2.5 Non-verbal Behavior Interpretation

In the previous Section, we reported the studies describing how specific non-verbal behaviors can bear information about attitudes. The interpretation of non-verbal behaviors depends on the context in which they are produced. For example a smile produced while our interlocutor is upset will have a very different interpretation than if it is produced in response to our interlocutor's smile. The meaning of a behavior clearly depends on the behavior produced by the other persons involved in the interaction. Similarly, a person's smile will be interpreted in a certain way if this person looks straight and smiles than if this person smiles then shakes her head and looks down. So the fact is that the perception of non-verbal signals is directly influenced by two key elements: sequentiality (order) and temporality (starting time and duration) of the surrounding non-verbal behaviors displayed by the interlocutors.

Sequentiality: Burgoon and Le Poire report that non-verbal signals can not be interpreted in an isolated way: *"what illuminates the interpretation of a given behavior is its accompanying composite of nonverbal cues. No nonverbal cue is an island. It is continually surrounded by a host of nonverbal behaviors which together may delimit and clarify meaning"* [Burgoon, J. K. and Le Poire, 1999]. Thus, in order to correctly interpret a non-verbal signal we should consider its context defined by its surrounding signals. For example, Heylen *et al.* showed that signals *tension* and *frown* do not mean "dislike" when displayed separately, whereas their combination does [Heylen et al., 2007]. In the context of interpersonal attitude, *averted gaze* has been reported as a sign of submission [Gifford, 1991, Debras and Cienki, 2012, Stivers, 2008, Bee et al., 2009]. However, this signal leads to an increase in the perception of dominance when it is followed by an expression

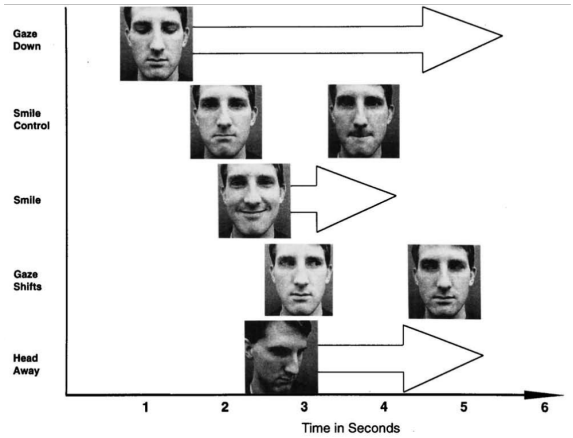


Figure 2.5 – Representation of a prototypical embarrassment response. The mean duration of each action is equal to the beginning of the interval with the leftmost edge of the photography and ending with the end of the arrow [Keltner, 1995].

of anger [Bee et al., 2009]. Based on these observations, sequential models have been developed to study how combinations or sequences of signals can reveal and influence the perceived emotions [With and Kaiser, 2011, Niewiadomski et al., 2011, Jack et al., 2014]. Results show that the expression of emotions as sequences of signals is better recognized than the static expression (one signal). Inspired by this work, Chollet *et al.* designed a virtual recruiter expressing interpersonal attitudes through sequences of non-verbal signals. A user study showed promising results of the proposed model. In this thesis, we adopt the representation of attitudes as multimodal sequences of non-verbal signals.

Temporality: non-verbal behaviors are also characterized by two other factors: starting time and duration. Keltner demonstrated that the starting time and duration of *smile*, *gaze shift* and *head away* can help differentiating between the possible meanings such as embarrassment, amusement and shame. Figure 2.5 represents the starting time and duration of some signals related to the expression of embarrassment [Keltner, 1995]. We can observe that although *smile* is usually considered as a cue of friendliness, it reflects more likely embarrassment if it is followed by a *gaze shift*. Signal duration is also important for behavior perception. For example, the duration of a smile could differentiate between faked and genuine smiles [Ekman and Friesen, 1982, Mcdaniel and Si, 2014, Keltner, 1995]. In the study of Keltner the mean duration of smile is 2.23 seconds whereas the mean duration of controlled (fake) smile is 0.46 seconds.

To sum up, the interpretation of a non-verbal behavior is influenced by three crucial factors: order, starting time, and duration. Our goal is to design an attitude model that encompasses all these three aspects: order, starting time and duration of non-verbal signals. As it will be detailed in Chapter 4, we rely on temporal sequence mining algorithm that best fits the modeling requirements.

This chapter laid the theoretical bases for attitude representation, measurement, and interpretation. The literature overview underlines that attitudes are interpersonal, multimodal and dynamic. Concerning the representation of attitudes, different affective di-

mensions can be used such as dominance and arousal. The interpersonal circumplex (IPC) stands out as the most popular representation of attitudes in the field of virtual agents. In this chapter, we have also described IPC inventories and two statistical approaches (circular profile and vector scoring) used to interpret these inventories.

On another hand, the previous works underlined that human behaviors are naturally multimodal and sequential: we interact with each other through multiple communication channels (speech, gaze, gesture, etc.). Moreover, these behaviors are temporally coordinated. Our goal is to understand how those behaviors are coordinated at critical moments, the sequential patterns they exhibit and their association with different interpersonal attitude variations. Thus, in our work, we choose to represent attitude variations as *multimodal* and *temporal* sequences of non-verbal signals.

Take home

- Interpersonal attitudes are multimodal and dynamic. Attitudes are expressed through verbal and non-verbal behaviors.
- The interpersonal circumplex (IPC) is the common representation of attitudes. IPC is composed of two orthogonal dimensions: friendliness (ranging from hostile to friendly) and dominance (ranging from submissive to dominant).
- Interpersonal attitudes are expressed through sequences of: non-verbal behaviors such as head movements, postures and facial expressions.
- The interpretation of non-verbal signal is influenced by its surrounding signals, their starting times and durations.

Chapter 3

State of the Art on Attitude Modeling for Virtual Agents

Contents

3.1 Attitude Modeling for Embodied Conversational Agents **21**
3.1.1 ECA's Behavior Expressing Attitude 22
3.1.2 Attitude Dynamics over Time 24
3.1.3 Attitude Generation Models 25
3.2 Sequence-Based Multimodal Behavior Modeling **26**

EMBODIED Conversational Agents (ECAs) are widely used in different fields. They are increasingly becoming essential elements by playing key roles: tutor, doctor, recruiter, etc. The general topic of this thesis is to conceive ECAs able to change their attitude toward the user according to their role as well as the context of the interaction. For example, it should be dominant with a job applicant and friendly with an autistic child. In this chapter, we present an overview of the most relevant works related to our topic: attitude modeling for virtual agents. We also focus on the works relying on sequentiality and temporality of non-verbal behavior as key components for human or agent behavior modeling.

3.1 Attitude Modeling for Embodied Conversational Agents

Existing works that model ECA's attitudes address different questions: which ECA's behaviors are the most influencing on the perception of its attitude? how does ECA's attitude change over time? how to automatically generate the ECA's behavior depending on its interpersonal attitude?

Figure 3.1 – Examples of an ECA showing non-verbal behavior (control: no behavior, dominant: akimbo (hands on the hip), submissive: neck-adaptor, cooperative: head tilt right and non-cooperative: gaze aversion) [Straßmann et al., 2016].

3.1.1 ECA's Behavior Expressing Attitude

Several investigations have been conducted to understand the impact of non-verbal behavior on attitude perception during human-human interaction (cf. Chapter 3). A lot of works relied on these studies for the selection of relevant non-verbal signals that potentially express attitudes and simulated them into virtual agents.

To begin with, Bee et al. studied the impact of the facial expression, gaze and head direction on the perception of virtual agent's dominance [Bee et al., 2009]. Results replicate some findings concerning the relationship between attitudes and the studied behavior from human-human interaction. For example, the expression of joy, anger and disgust have been strongly correlated with dominance perception compared to fear and sadness expressions [Knutson, 1996, Tiedens et al., 2000, Carney et al., 2005]. Moreover, an ECA with a bowed head is perceived as submissive while a raised head expresses dominance [Gifford, 1991, Debras and Cienki, 2012, Stivers, 2008]. Unlike what is generally reported in literature from human-human interaction, averted gaze of the agent did not influence the perception of submission. Later on, this study has been completed by adding linguistic behaviors to facial expression, gaze and head direction in order to investigate which modalities contributes the most to the expression of dominance [Bee et al.,]. The linguistic behavior integrates personality traits (agreeableness, extraversion and introversion) and the gaze model includes two states: looking at the user and looking away from the user. The drawn analyses suggests that both verbal and the nonverbal channels participate equivalently to the expression of dominance.

Straßmann *et al.* explored the perception of a virtual agent expressing dominance, submission, or cooperativity [Straßmann et al., 2016]. Based on the literature, they collected 23 non-verbal signals that are assumed to evoke those three attitudes. These behaviors are simulated using a virtual agent and evaluated through a perceptive study (see Figure 3.1). The obtained results reveal that gestures such as crossing the arms and laying the hands on the hip significantly influence the perception of dominance. However, for cooperativity, facial expressions have the most pronounced effect.

Other works investigate the interplay between attitudes and others behaviors such as interruptions and attention guiding. Cafaro *et al.* investigate how the interpersonal atti-

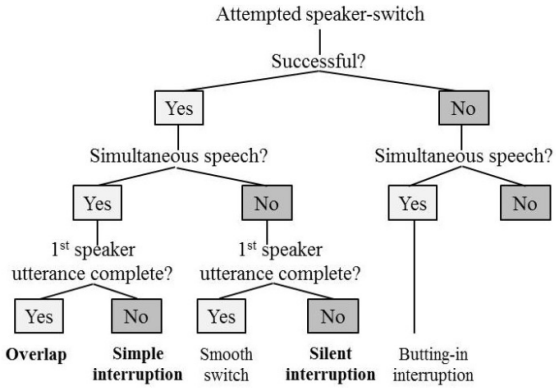


Figure 3.2 – Beattie's taxonomy of interruption types (image from [Cafaro et al., 2016b]).

tude (hostility/friendliness) and the personality (extraversion) of a virtual agent influence the first impressions of the users about the agent [Cafaro et al., 2012]. Based on literature, they consider three non-verbal signals: smile, gaze and proximity. For evaluating user's impressions about a virtual agent displaying a combination of these behaviors, they conducted a perceptive study in which users could interact in a virtual scene populated with virtual agents through their avatar. The authors found that users formed impressions of the agents after only 12.5 seconds of the beginning interaction. Proximity influenced the evaluation of extraversion: the agents approaching the subject's avatar were judged as more extraverted than those not approaching. However, smile and gaze were linked to friendliness expression. This work is limited to the context of first impressions and the expression of friendliness.

Cafaro *et al.* studied the influence of interruption types (amount of overlap between speakers and utterance completeness) on the perception of interpersonal attitudes during an agent-agent interaction [Cafaro et al., 2016a]. They designed three interruption types following the Beattie's taxonomy indicated in Figure 3.2: overlap, simple interruption, and silent interruption [Beattie, 1981]. Results revealed that the interruption types directly influence the perception of attitudes of both agents (interruptee and interrupter): the interruptee is perceived more dominant (and less friendly) when the amount of overlap increases. This work only considers interruption as indicator of attitudes.

Rosenthal *et al.* studied the effects of dominant and submissive non-verbal behavior of a virtual agent on the user perception of the agent in term of dominance, likeability, competence, autonomy, cooperativity, and communicative abilities [Rosenthal-von der Pütten et al., 2019]. They also explored attention guiding behavior (e.g., deictic gestures and gaze) and its impact on the perception of the dominance behavior. Based on previous works, they assumed that attention guiding behavior, in combination with dominant behavior, increases the perception of dominance. They categorized behaviors as follows:

- Dominant behaviors: akimbo posture, crossing arms, head up, gesture with large radius.

- Submissive behaviors: neck-adapter (self-touch), arms open, head down, gesture with small radius.

- Attention guiding: looking at the objects, looking and pointing at the object that is currently described when referring to it and looking back to the user after completing its utterance.

In order to explore the bias of the age difference, they consider two groups of participants: young adults and seniors. For both groups, the analysis reports that the virtual agent showing dominant behavior is perceived as more dominant than a virtual agent showing submissive behavior. However, there are no effects of nonverbal dominance on the perception of likeability, competence, and cooperativity. The attention guiding behavior increased the perception of dominance. Regarding age differences, seniors rated the agent as more likable, more autonomous, and more cooperative.

The works presented in this section focus on the impact of some behaviors on the perception of ECA's attitude. They do not extend to the dynamics and evolution of attitudes over time, which would be of a great importance for a comprehensive understanding of the topic. The following section will shed light on this aspect.

3.1.2 Attitude Dynamics over Time

Several works focused on the dynamics of attitudes by modeling the evolution of an ECA's attitude over time [Kasap et al., 2009, Ochs et al., 2010, Pecune et al., 2016]. Usually, the attitudes are first initialized w.r.t. the role of the agent and of its interlocutor. For example, an ECA assuming the role of a policeman will be initialized with a high dominance, respectively low dominance, value when interacting with a gangster, respectively with his superior. Then, depending on the emotion felt by the agent, its attitude will be adjusted. The models of this type are based on the Ortony, Clore AND Collin's (OCC) theory of emotion to formalize the emotion felt according to pre-defined rules [Ortony and Clore, 1988]. For example a virtual teacher can feel a gloating emotion when his difficult student got a bad result.

In this vein of works, Kasap and colleagues [Kasap et al., 2009] developed Eva, a virtual teacher that interacts with its students over several sessions. At the beginning of each session, the attitude of Eva is initialized according to its social attitudes achieved at the end of the previous session. Then, Eva's attitude is updated at the end of each session according to the emotions it felt during the current session. Positive emotions of gratitude and joy increase (resp. decrease) Eva's friendliness (resp. dominance) whereas the negative emotions of anger and distress decrease (resp. increase) its friendliness (resp. dominance). Eva has been evaluated in interaction with two students: the first one played a role of good student, whereas the second played the role of difficult student. During the first session with the first student, Eva was calm and polite and responded accordingly. The second student was rude, so Eva was more aloof in its responses. In the next sessions, Eva remembered each student's attitude and responded accordingly.

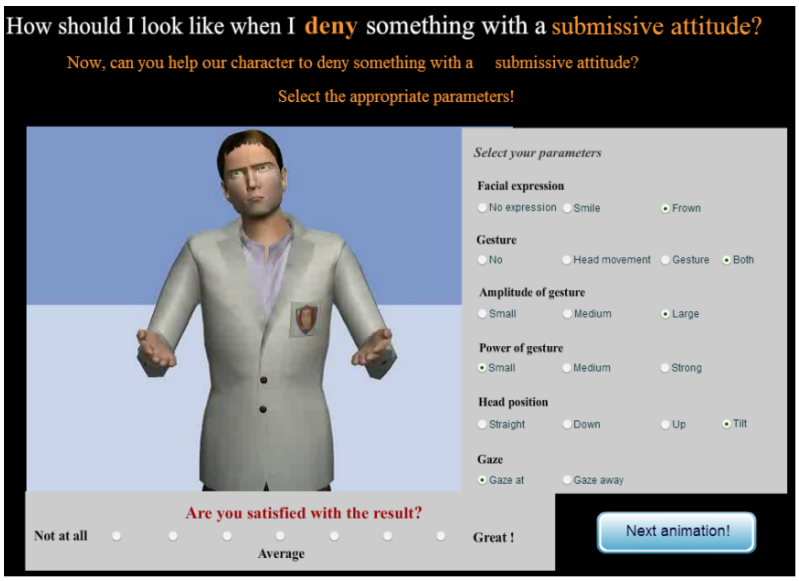


Figure 3.3 – Screenshot of the online graphical application designed to annotate the behavior of the ECA for a given attitude [Ravenet et al., 2013].

3.1.3 Attitude Generation Models

Several models of attitude have been designed in order to generate the behavior of a virtual agents. To learn the mapping between attitudes and non-verbal behavior, a corpus of ECA's non-verbal behavior conveying attitudes has been gathered and annotated using crowdsourcing [Ravenet et al., 2013]. For each attitude, the annotators indicated the facial expressions, gestures, head movement and gaze of the ECA used to express specific attitudes (see. Figure 3.3). Then, a Bayesian model has been designed in order to automatically generate the non-verbal behavior of an ECA given as input its interpersonal attitude. The Bayesian model has been combined with a dialogue model to integrate the verbal expression of attitude. A perceptive study was conducted with three conditions: the agent displayed an attitude through verbal behavior only, non-verbal behavior only, and both verbal and non-verbal behaviors [Callejas et al., 2014]. Results showed that only friendliness has been correctly perceived when the agent displayed only its attitude through the non-verbal behaviors (using the Bayesian model). Regarding the the verbal condition (dialogue model) no attitude has been recognized. However, the combination of both models lead to better perception of friendliness and hostility. Finally, both models have been used for the generation of the agent's behavior supporting a group interaction and the expression of friendliness and hostility [Cafaro et al., 2016b]. Despite the significance of this work, it only considered the friendliness dimension of attitude. In addition, it did not consider the behavior of the user when modeling the attitude of the agent.

In [Chollet et al., 2014b], using a corpus of job interviews between human recruiter and human job seeker, GSP algorithm has been applied to extract non-verbal sequences of a recruiter when s/he expresses different interpersonal attitudes toward a candidate. Then, an attitude planner have been developed to generate the behavior of the agent ac-

cording to its attitude and its communicative intention. This model was evaluated through perceptive studies, either using third-party protocol (where participants evaluated videos of an agent) or having participants interact with a virtual character. Most attitudes of the agent were recognized. They were better recognized in the third-party protocol than in the interactive study. A possible explanation support that the agent displayed an attitude only when it talks, not when it listens. The main contribution of this model is to consider both attitude and communicative intentions of the agent at time. The main limitation of this works is that the behavior of the candidate was completely ignored when modeling the attitude of the recruiter. Moreover, this work do not consider the temporal information (starting time and duration) of non-verbal behavior.

Table 3.1 recapitulates the presented works of attitude modeling. Each one is characterized with its goal that can be either (1) investigation: studying which ECA's behaviors are the most influencing on the perception of its attitude, (2) attitude dynamic modeling, and (3) generation: generating ECA's attitude. Moreover, for each study, we indicate how attitudes are represented as well as the considered dimensions of attitudes (dominance, friendliness, or both). Finally, we indicate how the attitudes are expressed or computed (based on verbal, non-verbal behaviors, or both).

Ref.	*Goal*	*Represent.*	*Attitudes*	*Behavior*
[Bee et al., 2009]	investigation	-	dominance	non-verbal
[Straßmann et al., 2016]	investigation	-	dominance	non-verbal
[Cafaro et al., 2012]	relationship	Argyle	both	non-verbal
[Cafaro et al., 2016a]	relationship	Argyle	both	non-verbal
[Kasap et al., 2009]	dynamics	Argyle	both	verbal
[Pecune et al., 2016]	dynamics	Argyle	both	verbal
[Chollet et al., 2014b]	generation	Argyle	both	verbal, non-verbal
[Cafaro et al., 2016b]	generation	Argyle	friendliness	verbal, non-verbal

Table 3.1 – A comparison between attitude models.

As underlined from this overview, attitudes are expressed through verbal and non-verbal behaviors. Thus, combining both modalities leads to better attitude recognition [Bee et al., , Callejas et al., 2014, Chollet et al., 2017]. However, only a couple of works did this combination [Chollet et al., 2014b, Cafaro et al., 2016b]. The rest of works used the representation of Argyle and considered the two attitude dimensions. Most of the presented models rely on non-verbal behavior to express a given attitude. However, no one leverages the temporal information of these behaviors. Our work addresses this limitation by considering the temporal information of non-verbal behaviors.

3.2 Sequence-Based Multimodal Behavior Modeling

Some researchers pointed out the importance of other characteristics to better model attitudes such as the sequentiality and temporality of non-verbal behaviors. As mentioned in Section 2.5, considering sequentiality and temporality (staring time and duration) of

non-vernal signals is determinant for interpreting the meaning of these signals. In the next section, we present existing works that encompass the sequentiality of non-verbal behavior in order to understand and predict phenomena such as emotion and interpersonal attitude.

Niewiadomski *et al.* [Niewiadomski et al., 2011] proposed a constraint-based approach to generate expressions of emotions for virtual agents. Their model includes: (i) a multimodal set of behaviors, extracted from both literature and annotated corpora; (ii) a set of spatial and temporal constraints in the form of hand-crafted rules were determined. These rules describe the temporal (i.e., the order and timing of behaviors) and the spatial (i.e., the multimodality) relationships regulating behaviors. However, the approach has some limitations. The corpora that have been used is small and the need for manual work to establish the rules makes the task costly and labor intensive.

Lately, sequence-mining algorithms like GSP, have been used to the task of extracting sequences of behaviors. Chollet *et al.* used GSP algorithm to extract, from a job interview corpus, non-verbal sequences representing interpersonal attitudes of a recruiter [Chollet et al., 2014b]. For example, his model extracted the sequence: head nod followed by smile for the expression of friendliness.

In [Martínez and Yannakakis, 2011], GSP algorithm has been used to discover which frequent multimodal sequences predict the best the emotional states of participants. The algorithm was applied to a game survey dataset and relied on three modalities: physiological signals (e.g. blood volume pulse), context-based game metrics (e.g. keyboard presses) and affective preferences. The obtained sequences have been transformed into feature vectors and presented as inputs to an Artificial Neural Network (ANN), specifically trained to predict affective states of players. A comparison between sequential and statistical features shows accuracy improvement for predicting players' affects when using sequential features. Despite this positive result, this type of modeling is limited by the intrinsic algorithm (GSP) which is not time-aware; it only considers the order of signals, but neither their duration nor their time of occurrence.

Other non time-aware algorithms have been used like T-Patterns [Magnusson, 2000]. With *et al.* automatically extracted sequences of facial expressions characterizing emotions [With and Kaiser, 2011]. T-Patterns algorithm has been used to detect sequences of facial signals representing five emotions: enjoyment, hostility, embarrassment, surprise, and sadness. A model of participation styles in collaborative learning interaction has also been proposed in [Nakano, 2015] using the multidimensional motif discovery algorithm [Vahdatpour et al., 2005]. As such, 122 behavioral patterns of participants have been extracted. One example pattern indicated that, while the target expert participant was speaking, the other expert participant and the novice participant gazed both at the target participant. This model only focuses on gazing and turn-taking of participants.

Zhao *et al.* used TITARL algorithm [Guillame-Bert and Crowley, 2012] to predict behavioral patterns that convey a variation in interpersonal rapport [Zhao et al., 2016]. TITARL allows predicting temporal relation between signals like occurrence interval (e.g., if there is an event d at time t, then there is an event c at time $t + 5$). For interpersonal rapport modeling, a corpus involving a tutor and tutees has been annotated on several

levels: gaze, smiles, conversational strategies like social norm violation, and interpersonal rapport. TITARL algorithm has been applied to extract temporal association rules representing either an increase or a decrease of rapport between tutor and tutees. For example, *the tutor violates social norms while being gazed at by the tutee, and their speech overlaps within the next minute*. The TITARL algorithm has also been used in [Janssoone, 2016] to extract temporal association rules related to attitude from the SEMAINE database [McKeown et al., 2012]. More precisely, the authors investigated the correlation between non-verbal behavior (like eyebrow movements and prosody), and two attitudes: friendliness and hostility. TITARL allows predicting temporal relation between signals. However, it does not extract exact duration of signals.

The works in [Fricker et al., 2011, Yu et al., 2010, Zhang et al., 2010] have focused on extracting temporal sequences of non-verbal behaviors from human-robot interactions. The extracted information has been used to analyze human's behavior, primarily gaze behavior, in relation to the robot's behavior. In addition to explicitly considering timing of non-verbal behaviors, these works are set in dyadic settings; i.e., they consider both human's and robot's behavior. However, they are not generative; the extracted patterns are not explored for generating robot's behavior.

Finally, probabilistic models like Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) have also been used to predict appropriate sequences of multimodal signals in human interactions. The goal is to predict a given behavior (backchannels, gestures) for each time window t based on a sequence of multimodal signals (features) observed over t. Lee and Marsella predicted speaker's head movements using HMM models [Lee and Marsella, 2010]. Different feature types have been considered: phrase boundaries, part-of-speech tags, dialog act, etc. In [Lee and Marsella, 2012], the authors used HMM and CRF to estimate head nods and eyebrow movements of the speaker.

The Dynamic Bayesian Networks (DBNs) and Deep Conditional Neural Fields (DCNF) have also been used for social behavior modeling. For example, a DBN has been designed to estimate turn taking [Otsuka et al., 2007] or to predict gaze and hand gestures of the instructor in a collaborative task [Mihoub et al., 2016]. Chiu *et al.* experimented deep conditional neural fields to model the generation of gestures by integrating verbal and acoustic modalities [Chiu et al., 2015]. These kind of methods can not provide the exact starting time and duration of signals.

As shown above, the integration of sequentiality into behavior modeling has been approached from various perspectives: sequence mining, probabilistic modeling, etc. Table 3.2 gives a comparison of the works presented above. Each one is characterized with a set of criteria: underlying algorithm used to automatically extract behavioral sequences, consideration of signal timing (start time and duration), mono or multimodality, and generation (implementation of the extracted patterns into virtual agents). Also, we indicate the goal behind each work (e.g., attitude modeling, emotion modeling, etc.)

As one can see, most of works only rely on the order of signals ignoring their temporal information (i.e., their starting time and duration [Chollet et al., 2014b, Lee and

Ref.	*Algorithm*	*Signal timing*	*Modality*	*Gener--ation*	*Goal*
[Niewiadomski et al., 2011]	None (manual)	✓	multimodal	✓	emotion modeling
[With and Kaiser, 2011]	T-Patterns	x	facial, head	x	emotion
[Fricker et al., 2011]	ESM	✓	gaze, head	x	HRI analysis
[Zhang and Boyles, 2013]	QTempIntMiner	✓	facial, head	x	HRI analysis
[Lee and Marsella, 2010]	HMM	x	head nod	✓	head movements
[Lee and Marsella, 2012]	LDCRF	x	face	x	head movements
[Chollet et al., 2014b]	GSP	x	multimodal	✓	attitude
[Janssoone, 2016]	TITARL	partially	face		attitude
[Zhao et al., 2016]	TITARL	partially	multimodal		rapport

Table 3.2 – A comparison between sequence-based behavior modeling methods.

Marsella, 2012, Lee and Marsella, 2010, With and Kaiser, 2011, With and Kaiser, 2011]). Some works consider a limited number of modalities [Fricker et al., 2011, With and Kaiser, 2011, Yu et al., 2010, Zhang and Boyles, 2013], while others rely on hand-crafted constraints [Niewiadomski et al., 2011]. Only a couple of these works explored the extracted sequences of human behaviors for generating virtual character's behaviors [Chollet et al., 2014a]. Our work addresses all these limitations by considering the temporal information in human behaviors. We propose a fully-automatic, sequential, temporal and generative model for extracting non-verbal sequences representing different attitude variations.

Take home

- Different models of attitudes have been developed for different purposes such as attitude dynamics modeling and attitude behavior generation. Most of them rely on the non-verbal behavior for attitude expression and adopt the Argyle's model for attitude representation.
- We presented existing works that model the sequentiality of multimodal behaviors. Most of previous works only rely on the order of signals ignoring their temporal information. Moreover, most of them are not generative.

Chapter 4

Sequence Mining: State of the Art and Our Algorithm

Contents

4.1 Non-Temporal Algorithms **32**
4.2 Temporal Algorithms **33**
4.3 Temporal Sequence Mining Algorithms **35**
4.4 HCApriori Algorithm **37**
4.4.1 Definitions 38
4.4.2 Algorithm 39
4.4.3 Evaluation and Results 41
4.5 Pattern Quality Assessment **43**
4.6 Conclusion **45**

IN our work, we represent interpersonal attitudes as sequences of non-verbal signals. Relying on sequence mining algorithm, we extract, from a multimodal corpus, the most relevant sequences of behaviors characterizing attitude variation. Sequence mining is a data mining task that aims at discovering relevant patterns hidden in a set of sequences. A pattern is a sub-sequence that occurs frequently in the dataset. Sequence mining has been applied in a wide range of real-life applications in many domains such as market basket analysis, text mining, bioinformatics, and human behavior analysis [Chollet et al., 2017, Fricker et al., 2011]. For example, in the context of market basket analysis, sequence mining can be used to identify the sequences of items frequently bought by customers. This can be useful to understand the behavior of customers to take marketing decisions.

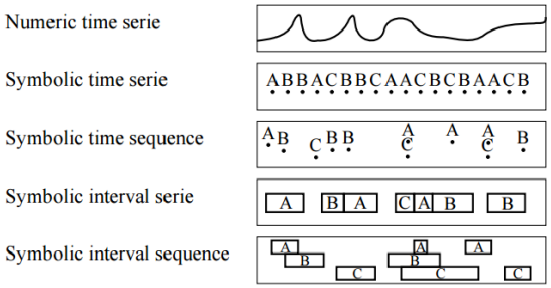


Figure 4.1 – Temporal data model.

Various sequence mining algorithms have been developed to explore different temporal information such as order, starting time, duration, etc. Different temporal data models have been introduced in [Mörchen, 2007] and summarized in Figure 4.1. For example, in biology, symbolic time series are used to represent DNA fragments. However, non-verbal signals can only be represented by symbolic intervals sequence since communication is multimodal (more than one non-verbal signal can occur simultaneously) and temporal (a signal appears at a given time and for a given duration). Then, order-based algorithms are adapted for DNA analysis but not for non-verbal behaviors modeling because it requires a temporal algorithm that takes into account starting time and duration of signals.

In this chapter, we present the different algorithms of sequence mining divided in two categories: no temporal algorithms (Section 4.1) and temporal algorithms (Section 4.2). In Section 4.4, we introduce a new temporal sequence mining algorithm HCApriori for Hierarchical Clustering Apriori to overcome the limitations of existing algorithms. Finally, in Section 4.5, we present the metrics that are commonly used to assess pattern quality and that are based on occurrence frequency. We enhance these metrics by considering signal temporality.

4.1 Non-Temporal Algorithms

Apriori algorithm was designed to discover patterns in transactions made by customers in stores [Rakesh Agrawal, 1994]. The goal is to better understand the behavior of customers and to take future marketing decisions. For example, 10% of customers bought butter and eggs at the same time. Apriori takes as input a minimum frequency threshold (f_{min}) and a transaction database containing a set of transactions. A transaction is defined as a set of distinct items (cf. Table 4.1). Apriori outputs all frequent itemsets, i.e. set of items that occur in more than f_{min} transactions in the input database.

The Apriori algorithm is made of two phases: (1) join phase in which frequent itemsets of size n are extended to generate candidate itemsets of size $n+1$. For example, from the items $\{A, B, C\}$, it generates the following itemsets candidates of size 2: $\{A, B\}$, $\{A, C\}$ and $\{B, C\}$. (2) Prune phase: all the itemsets that occur less than f_{min} are deleted from the candidate sets. Join and prune steps are performed repetitively until no more patterns can be generated. For example, using the transaction database indicated in Table 4.1 and

Transaction 1	A, B, C
Transaction 2	B, D
Transaction 3	A, E, B
Transaction 4	F, A, E

Table 4.1 – Transaction database.

$f_{min} = 2$ transactions, frequent itemsets of size 1 are A, B and E. A and B appear in 75% of the dataset, whereas E only occurs in 50% of the dataset. We also find two itemsets of size 2 A, B and A, E. Both items occur in 50% of the dataset.

Apriori was initially designed to better understand customer behavior but it was also applied in wide range of applications from text analysis to medical data analysis. For example, Apriori was used for analyzing university admission data [Mashat et al., 2013]. One extracted information is that 62.7% of the rejected students are females and studied literature study in high school. Srikant extended Apriori by considering event order which gave rise to the first sequence mining algorithm GSP [Srikant and Agrawal, 1996]. Taking as input a sequence dataset and f_{min}, GSP discovers frequent patterns (sub-sequences) based on simple ordering of signals. For example, from the dataset $\{ ABB, ABC, CABA, CABCA\}$ with $f_{min} = 2$, the frequent patterns of length 3 are $\{CAA, ABA, ABC, CAB\}$. However, relying only on event's order may become a limitation where timing such as starting time, duration, and delay between events are important information.

4.2 Temporal Algorithms

Temporal algorithms are designed to address the time-related issues such as: what is the delay between two temporal events? At what moment a temporal event happens? And what is its duration? The existing time-aware methods can be regrouped into three categories:

1. Relation-based models: they are deployed along with Allen's interval relations (meets, overlaps, starts, before, during, ...) to detect symbolic temporal relations between events [Kam and Fu, 2000, Höppner, 2002]. The Allen's relationships are shown in Figure 4.2. In addition to order, the events are related to each other with temporal relations. For example, ((A overlaps B) before C) overlaps D. The extracted patterns could have several representations because the same Allen's relation can represent very different situations. For example, the temporal sequence A(6, 12), B(9, 14), D(17, 24), C(20, 22) can be represented by two different patterns as indicated in Figure 4.3;

2. Interval-based models: they focus on the interval of occurrence and the delays between events [Nakagaito et al., 2009, Guillame-Bert and Crowley, 2012, Chen et al., 2003]. For example, A appears between 5 and 10 sec. These algorithms do not provide exact starting time and duration of events but attribute them to intervals;

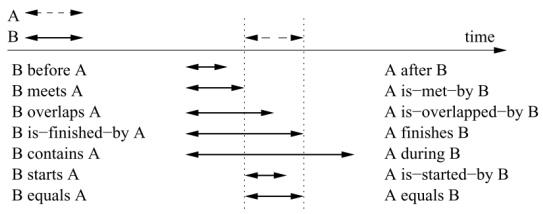


Figure 4.2 – Allen's interval relationships [Kam and Fu, 2000].

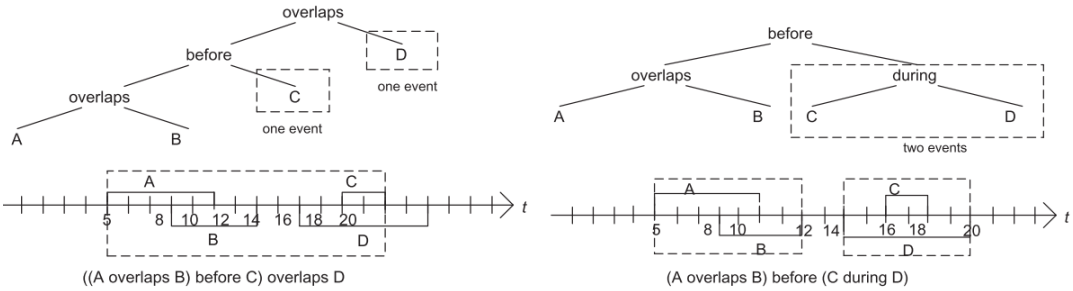


Figure 4.3 – Different representations of a temporal pattern [Kam and Fu, 2000].

3. Temporal sequence mining algorithms: they extract the exact timing of events [Guyet and Quiniou, 2008, Ruan et al., 2014]. For example, A from sec. 3 to sec. 5 followed by B from sec. 6 to sec. 8. This last category leads to more informative patterns that can be used to represent the whole temporal information given by the two other categories.

In our work, we focus on temporal sequence mining algorithms as they allow answering our research questions: given the current context (defined by the previously occurring non-verbal signals), (i) at what moment a signal must happen? And (ii) what is its duration? For exact timing extraction, these methods combine classical sequence-mining algorithms, usually Apriori, with a data clustering algorithm as illustrated in Figure 4.4. For example, events can be projected in 2-D space formed by the axes "starting time" and "duration" (Figure 4.4.1). Then, a clustering algorithm allows grouping events that mostly occur at the same time (Figure 4.4.1). The centroid of each cluster will represent one temporal pattern of size 1 (Figure 4.4.2). Finally, Apriori-like procedure will be applied repetitively until no more patterns can be generated (Figure 4.4.2). In addition to f_{min}, temporal sequence mining algorithms require two parameters to measure the temporal distance between events: a temporal dissimilarity measure used to evaluate the temporal distance between events; and a dissimilarity threshold (ϵ) that is used to decide if two events are temporally similar or not.

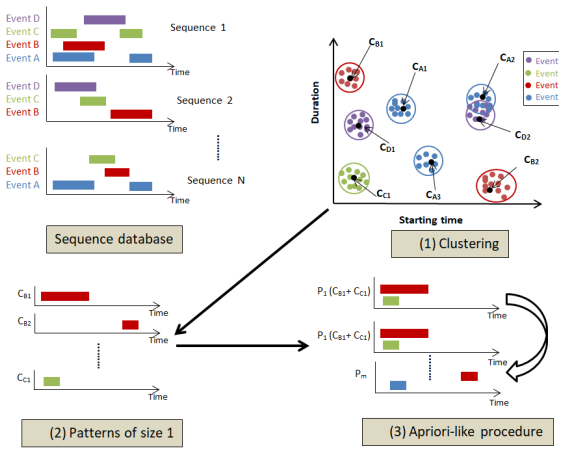


Figure 4.4 – Process of temporal sequence mining algorithms.

4.3 Temporal Sequence Mining Algorithms

In the following, we present existing temporal sequence algorithms. QTempIntMiner represents temporal sequences as hyper-cubes, with one dimension for each event type [Guyet and Quiniou, 2008]. Then, it uses Apriori simultaneously with a Gaussian mixture model for approximating the time distribution of the frequent patterns given by Apriori. However, QTempIntMiner has significant time complexity as it applies repeatedly clustering during the Apriori procedure. QTIPrefixSpan can be viewed as an extension of QTempIntMiner [Guyet and Quiniou, 2011]. It combines PrefixSpan and Kmeans or AP (Affinity Propagation clustering) [Pei et al., 2001]. Unlike Apriori that scans the entire database to generate a candidate pattern, PrefixSpan reduces the research space by eliminating the sequences that are not frequent in the previous iteration. The advantage of QTIPrefixSpan is to use more efficient algorithm (PrefixSpan), instead of Apriori, which globally reduces its complexity.

PESMiner follows a user-in-the-loop approach. It first implements the events to the user in 2-D space formed by the axes “starting time" and “duration" so that a user can manually choose the cluster centroids for each event type [Ruan et al., 2014]. Then, a clustering step can be added to smooth the manually-defined centroids that have been selected. In PESMiner, clustering is performed once for each event type. The duration of patterns are adjusted based on a Gaussian distribution of the event’s duration. The main limit of PESMiner lies in its semi-automatic nature (the user must choose the initial candidate patterns (the cluster centroids)). The output of such algorithms depends highly on the initial candidate patterns. Thus, the first selection is a crucial step.

The algorithm that we propose, HCApriori, overcomes the main limitations of the existing algorithms: it is fully automatic and it increases cluster homogeneity by implementing an adapted (hierarchical) clustering technique. The motivation behind the development of HCApriori are summarized in the following points:

- **Better handling of sparsity in data**: sparsity is an important aspect to take into account when clustering data. Existing works did not focus on this phenomenon because they all have been applied on syntactic data, often not exposed to this problem (cf. Figure 4.5);

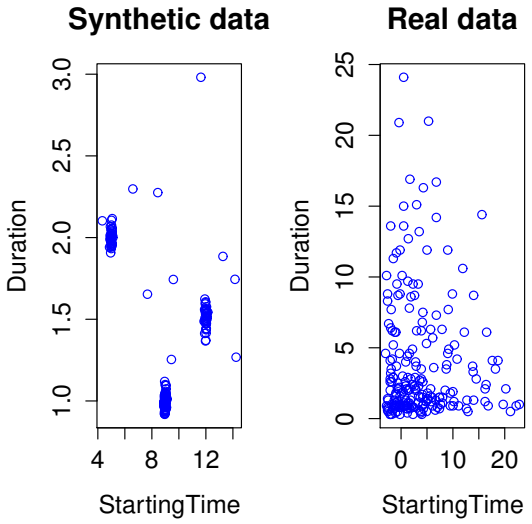


Figure 4.5 – Synthetic data vs. real data representing gaze signals. Real data is very sparse while synthetic data can easily be grouped into three clusters.

- **Better cluster homogeneity**: when using partitioning algorithms for clustering (like Kmeans), the events clustered together may be very distant in time. To overcome this limitation, our algorithm HCApriori relies on hierarchical clustering method that imposes a dissimilarity threshold to the events from the same cluster. This idea is widely used in data clustering for outlier detection using the so-called density-based methods [Breunig et al., 2000];

- **Fully automatic**: the existing algorithms require the user to provide an a priori estimation of the number of clusters or their approximate location. Thanks to hierarchical clustering, HCApriori does not require manual input. All the steps in the algorithm are fully automatic;

- **Customized and automatically-computed dissimilarity threshold**: in the previous existing works, the dissimilarity threshold ϵ is set by the user and it has the same value for all event types. However, this constraint can be restrictive as the event duration's may significantly differ w.r.t. event types. For example, the duration of a smile is likely to be much shorter than the duration of a posture as shown in Table 4.4. For more flexibility, HCApriori offers the ability to customize this parameter for each event type to the user. Moreover, setting manually the value of ϵ for each event type can be challenging because of the number of event types (27 in our corpus) present in the data and also because of the differences between the event types. To overcome this issue, we propose an efficient way to set up ϵ automatically.

Gaze At	Gaze Up	Eyebrow Up	Body Lean	Body Recline	Arms Crossed	Smile
4.46	1.26	2.34	32.34	17.41	11.74	2.01

Table 4.2 – Mean duration of some non-verbal signals in seconds.

Algorithm	Seq. Min. algorithm	Clustering	Dissim. meas. (δ)	Data	Dim.	Cluster hom.	Fully automatic	Cust. ϵ	Source code
QTempIntMiner	Apriori	EM	-	S	N-D	x	✓	x	Matlab
QPrefixSpan	PrefixSpan	QFIMiner	-	S	N-D	x	✓	x	Java
QTIPrefixSpan	PrefixSpan	Kmeans, AP	Haussdorf, CityBlock	S	N-D	x	✓	x	Matlab
PESMiner	Apriori	Kmeans	Euclidian	S	2-D	x	x	x	Java
HCApriori	Apriori	Hierarchical	Haussdorf, CityBlock	R	2-D	✓	✓	✓	Java

Table 4.3 – A comparison of temporal sequence mining algorithms.

- **Better time efficiency**: in [Guyet and Quiniou, 2008, Guyet and Quiniou, 2011, Nakagaito et al., 2009] the clustering is applied repeatedly in N-D space (one dimension for each event type to extract temporal patterns of length N) which increases the time complexity of these algorithms. The works of Ruan *et al.* [Ruan et al., 2014] inspired us to reduce the space dimension from N-D to 2-D ("starting time" $\times$ "duration"), which significantly reduces the time complexity of our algorithm.

Table 4.3 gives a comparison of the four algorithms presented above as well as our algorithm. Each algorithm is characterized with a set of criteria: underlying sequence-mining method, clustering technique, implemented dissimilarity measure, data type used for evaluation in the original paper (synthetic (S) or real (R)), space dimension where the clustering is applied (N-D vs. 2-D), cluster homogeneity, and customized ϵ.

AS shown in Table 4.3, HCApriori is fully automatic, Customized and automatically-computed the dissimilarity threshold ϵ. Unlike the four other algorithms, HCApriori rely on hierarchical clustering which allow better cluster homogeneity.

4.4 HCApriori Algorithm

To deal with the challenges of the sequence-mining algorithms highlighted in Section 4.3, we propose a new temporal sequence mining algorithm that we called HCApriori [Dermouche and Pelachaud, 2016b]. We have made our algorithm HCApriori opensource. It is available on github[1].

[1] https://github.com/dermosamo/HCApriori.git

4.4.1 Definitions

In this section, we give the formal definitions from the temporal sequence mining domain that are relevant to our problem (definitions 1–3). Next, we introduce new concepts on which we build our algorithm (definitions 4–5).

Definition 1 Temporal event
A temporal event e is a triplet (t, s, e), where e^t is the event type. e^s and e^e are the starting, respectively the ending, time of the event (with $e^s < e^e$). Consequently, the event duration $e^d = e^e - e^s$.

Definition 2 Temporal sequence
A temporal sequence S of length k is a sequence of temporal events $(e_1, e_2, \ldots, e_k)$ for which $\forall e_i \; for \; 1 \leq i < k : e_i^s \leq e_{i+1}^s$.
An event type can appear more than once at different times. Only events of different types can overlap. D denotes a set of temporal sequences and T the set of all events types in D. An event type can appear more than once at different times. Only events of different types can overlap.

$S = ((A, 2, 4.5), (B, 4, 8), (A, 5, 10))$ is a valid temporal sequence. All events are ordered according to their starting time and the two events of type A do not overlap.

Definition 3 Temporal event dissimilarity
Let e_1 and e_2 be two temporal events. A dissimilarity measure δ is a function that reflects the time difference between e_1 and e_2. For example:
$CityBlock(e_1, e_2) = |e_1^s - e_2^s| + |e_1^e - e_2^e|$
$CityBlock$ represents the sum of the time difference between the starting times and the ending times of e_1 and e_2.
$Haussdorf(e_1, e_2) = \max\{|e_1^s - e_2^s|, |e_1^e - e_2^e|\}$
$Haussdorf$ represents the maximum between the time difference of the starting times or of the ending times. Note that this distance is finite if and only if e_1 and e_2 have the same type. Otherwise, $\delta(e_1, e_2) = \infty$.

Definition 4 "matches" relation
Let ϵ be a given dissimilarity threshold. We define the binary relation "matches" that can be either between two events e_1 and e_2, between an event e and a temporal sequence S, or between two temporal sequences S_1 and S_2 as follows:

- *e_1 matches e_2 if $\delta(e_i, e_j) < \epsilon$.*

  In this case, e_1 can not match e_2 unless they are of the same type.

- *e matches S if $\min_u\{\delta(e, u)$ for $u \in S\} < \epsilon$.*

  We consider that the event e matches one event of the sequence S (for sake of simplicity, we write e matches S).

- *S_2 matches S_1 if $\forall e \in S_2 : e$ matches S_1.*

  Note that in this case, the relation is not symmetric.

Example 1: let $e_1 = (A, 2, 5)$ and $e_2 = (A, 3, 7)$ be two temporal events and $\epsilon = 2$. Based on the $CityBlock$ distance, we have:

$$\begin{aligned} \delta(e_1, e_2) &= |2 - 3| + |5 - 7| \\ &= 3. \end{aligned}$$

As it is greater than 2, e_1 does not match e_2.

Example 2: let $e = (A, 2, 5)$ be a temporal event, $S = ((A, 2, 4.5), (B, 4, 18), (A, 5, 10))$, and $\epsilon = 2$.
In this example, e matches S because the $min\{\delta((A, 2, 5), (A, 2, 4.5)), \delta((A, 2,), (A, 5, 10))\} = 0.5 < 2$.

Definition 5 Frequent temporal pattern
Let D be a temporal sequence dataset, $f_{min} \in [0, 1]$ is a fixed minimum support. We define a frequent temporal pattern P over D as being a temporal sequence that matches at least $f_{min} \times |D|$ sequences of D, that is:

$$|\{S \in D : P \text{ matches } S\}| \geq f_{min} \times |D|.$$

This definition will be used to check if a candidate pattern P is a frequent temporal pattern over D.

4.4.2 Algorithm

The novelty of our algorithm HCApriori is to customize the dissimilarity threshold (ϵ) for each event type (posture, gesture, gaze, etc.) and to propose an automatic computation of ϵ alternatively to manual setting. For outliers detection, HCApriori relies on hierarchical clustering that imposes a distance less than ϵ to the event from the same cluster. HCApriori operates in three steps: (1)first, the algorithm computes a dissimilarity threshold ϵ for each event type based on the duration of all events of same type. (2) Hierarchical clustering is applied to merge events into the same cluster if and only if their temporal distance is less than ϵ. At the end of this step, the cluster centroid represents a pattern of length one. (3) Taking as input the results of step (2), Apriori-like procedure is adapted to generate temporal patterns of length more than one. We now present each step in more details.

Step 1: dissimilarity threshold computation. The dissimilarity threshold ϵ represents the temporal distance between two temporal events e_1 and e_2 that is used to decide if they match or not (cf. Definition 4). As can be seen on the examples given in Table 4.4,

the event's duration may change w.r.t to the event's type. The novelty of HCApriori is to customize ϵ for each event type. The user has the possibility to give manually the value of ϵ for each event type. Otherwise, ϵ can be computed automatically as a function of the duration of all the events of the same type. E.g., the functions mean, quartile, percentile, can be applied. For example, based on the mean duration, ϵ for a given event type t is calculated as follows: $\epsilon_t = \text{mean } (e^d)$ for all $e^t = t$

Step 2: hierarchical clustering. Once the dissimilarity threshold ϵ is set for each event type t, we perform hierarchical agglomerative clustering on each event type separately. Initially, each event is considered as a single cluster on its own (cf. Algorithm 1, line 3). At each iteration, the clusters c_1 and c_2 with the minimum distance are merged (lines 6). The centroid of a cluster c is represented by the mean of all events in c.

Algorithm 1: HierarchicalClustering

Input : D, a temporal sequence database
n_{min}, the minimum number of events in a cluster
ϵ, dissimilarity threshold
Output: C, a set of centroids

```
1  C ← {};
2  foreach event type t in T do
3      clusters ← {e ∈ D where e^t = t} ;
4      {c1, c2} ← arg min_{ci,cj} δ(ci, cj) for ci, cj ∈ clusters and i ≠ j;
5      while δ(c1, c2) ≤ ε and |clusters| > 1 do
6          Merge the clusters c1, c2 into one;
7          {c1, c2} ← arg min_{ci,cj} δ(ci, cj) for ci, cj ∈ clusters and i ≠ j;
8      end
9      foreach cluster c in clusters do
10         if |c| > n_min then
11             C ← C ∪ {c.centroid};
12         end
13     end
14 end
15 return C
```

In addition to this, our clustering algorithm implements a cluster homogeneity criterion: c_1 and c_2 can not be merged unless $\delta(c_1, c_2) \leq \epsilon$. Otherwise, the clustering procedure converges and stops. We have added this criterion in order to isolate and ignore outlier events. Also, we choose to discard the clusters with less than a minimum number of events n_{min} (lines 9–13). By default, this parameter is set to 2.

Step 3: Apriori procedure. Based on Apriori algorithm [Rakesh Agrawal, 1994], our algorithm generates the frequent temporal patterns in two steps: a set of candidate temporal patterns of length $n+1$ is generated from all the temporal patterns of length n. Also, like the Apriori algorithm suggests, the infrequent patterns are pruned (cf. Definition 5).

Candidate generation and pruning are performed repetitively until no more patterns can be generated.

For candidate pattern generation, we adapt Apriori algorithm to take into account the temporal dimension of our data. Thus, to generate a new candidate pattern, a frequent temporal pattern of size 1, e, ($e \in L_1$, cf. Algorithm 2, lines 1–2) is appended at the end of a temporal pattern p if and only if the following two conditions are satisfied:

1. The starting time of e is greater than the starting time of the last event in p.
2. The starting time of e is greater than the ending time of the last event of type e^t from p;

The pseudo-code of HCApriori algorithm is given in Algorithm 2. Th code is implemented in JAVA and available on github[2] under GNU licence version 3.

Algorithm 2: HCApriori

Input : D, a temporal sequence database
f_{min}, minimum support
n_{min}, minimum number of events in a cluster
ϵ, dissimilarity threshold
Output: P, set of frequent temporal patterns

1 $C_1 \leftarrow HierarchicalClustering(D, n_{min})$;
2 $L_1 \leftarrow PruneInfrequentPatterns(D, f_{min}, C_1)$;
3 $n \leftarrow 1$; $P \leftarrow \{\}$;
4 **while** $|L_n| > 0$ **do**
5 $\quad P \leftarrow P \cup L_n$;
6 $\quad n \leftarrow n + 1$;
7 $\quad C_n \leftarrow CandidateGeneration(L_{n-1}, L_1)$;
8 $\quad L_n \leftarrow PruneInfrequentPatterns(D, f_{min}, C_n)$;
9 **end**
10 return P

4.4.3 Evaluation and Results

We evaluate our algorithm on a corpus of dyadic interactions where nonverbal behaviors and attitude variations of the dyadic interactions are annotated (cf. Section 5.1.1). We apply HCApriori to extract sequences of multimodal behaviors along with social attitude variations. We compare the results of our HCAapriori algorithm against the results obtained by the four state-of-the-art algorithms: QTIPrefixSpan-Kmeans, QTIPrefixSpan-AP, QTIApriori-Kmeans, and PESMiner. The comparison is based on pattern extraction accuracy criteria. The accuracy is defined as the percentage of sequences from the original data that are similar to at least one pattern from the set of extracted patterns. Two temporal sequences S_1 and S_2 are similar if the temporal distance between S_1 and S_2 is less than

[2] https://github.com/dermosamo/HCApriori.git

Gaze At	Gaze Up	Eyebrow Up	Body Lean	Body Recline	Arms Crossed	Smile
2	0.28	0.5	10.55	5.45	2.69	1.55

Table 4.4 – Example value of ϵ customized by event's type (non-verbal signals).

the dissimilarity threshold (ϵ).

$$\text{Accuracy} = \frac{|\{S \in D : \exists\, p \in P,\ p \text{ matches } S\}|}{|D|} \tag{4.1}$$

For the evaluation purpose, we rely on $CityBlock$ as a distance measure (see Definition 3) and we automatically set the threshold ϵ for each type t to 40% of the mean duration: $\epsilon_t = \text{mean}\ (e.d) \times 0.4$ for $e^t = t$. As such, we obtain values of ϵ that are different for every event type. Table 4.4 gives example values for some events types (non-verbal signals). This parameter is set to 1 sec. in the other four algorithms.

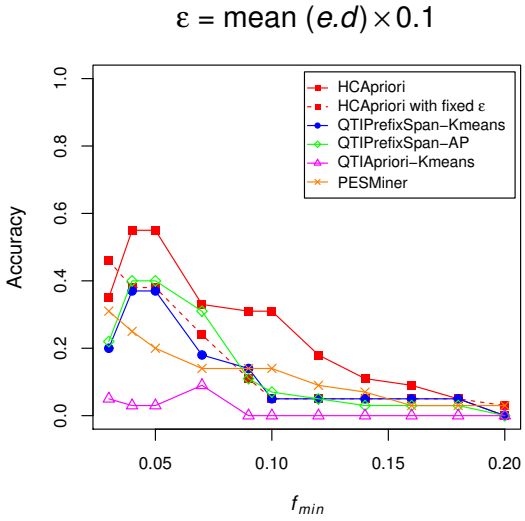


Figure 4.6 – Accuracies of the compared algorithms for different values of f_{min}. Here, ϵ is used for evaluation and its value is less than the one used for pattern extraction.

Figure 4.6, Figure 4.7 and Figure 4.8 plot the accuracy of the experimented algorithms as a function of f_{min} (minimum frequency threshold) for three different values of ϵ. As can be seen, the accuracy notably increases when ϵ increases: our algorithm achieves accuracy of 0.58, 0.82 and 0.92 for the three values of ϵ (10%, 20%, and 30% of the mean duration) respectively. Increasing accuracy is expected as ϵ directly controls the quantity of patterns that pass the minimum similarity filter. However, the accuracy notably decreases when f_{min} increases and becomes almost zero for $f_{min} \geq 0.2$. This is expected because the number of extracted patterns decreases when f_{min} increases. Concerning the comparison between the algorithms, we observe that our algorithm HCApriori outperforms the other algorithms and is able to achieve over 0.92 accuracy (Figure 4.8) whereas the runner-up

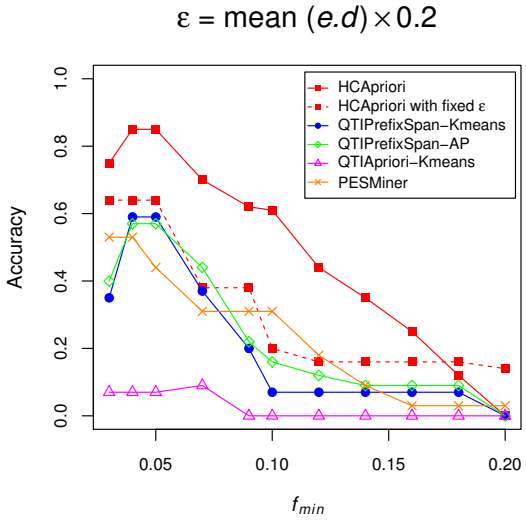


Figure 4.7 – Accuracies of the compared algorithms for different values of f_{min}. Here, ϵ is used for evaluation and its value is less than the one used for pattern extraction.

achieves 0.70. We believe the homogeneity criterion implemented by HCApriori is behind its good performance.

Moreover, in order to assess the efficiency of the customized setting of ϵ, we also measure the accuracy of HCApriori with a fixed value of ϵ (set to 1 sec. for all event types). The results of this experiment are represented in dashed lines in Figure 4.8. HCApriori algorithm generally achieves better performance (about 0.4 better) when using the customized setting of ϵ (customized by event's type). This observation validates the assumption that differences among events duration should be taken into account.

Figure 4.9 shows the running times of the five algorithms for different sizes of the dataset (f_{min} set to 0.05). We can observe that HCApriori and PESMiner run much faster than the other algorithms. While with the other algorithms, the running time rises linearly with the dataset size, HCApriori and PESMiner are hardly affected by the dataset size. This comes from the fact that both HCApriori and PESMiner applies clustering in 2-D space which greatly reduces its complexity.

In addition to this quantitative evaluation, we also perform some qualitative checks of the extracted patterns based on user study. The results of this study will be presented in Chapter 5.

4.5 Pattern Quality Assessment

Based on the occurrence frequency, the quality of an extracted pattern p is assessed using two quality measures: (1) $support(p)$ that indicates the frequency of the pattern p in the

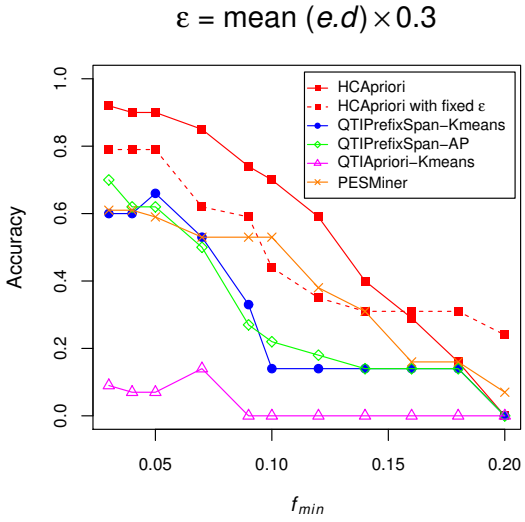


Figure 4.8 – Accuracies of the compared algorithms for different values of f_{min}. Here, ϵ is used for evaluation and its value is less than the one used for pattern extraction.

dataset D (eq. 4.2). (ii) $confidence(p, v)$, which reflects the proportion of D containing p and expressing the attitude variation v (eq. 4.3).

$$\text{Support (p)} = \frac{|\{S \in D : \text{S contains p}\}|}{|D|} \tag{4.2}$$

$$\text{Confidence (p, v)} = \frac{|\{S \in D : \text{S contains p and S expresses v}\}|}{|\{S \in D : \text{S contains p}\}|} \tag{4.3}$$

Despite their popularity, these measures present a major shortcoming regarding our domain: they did not consider the temporal relations between events. In order to provide a temporal similarity between the extracted patterns and the input dataset, we extend the classical measures $support(p)$ and $confidence(p, v)$ by considering the time overlap of signals. We define the overlap between two temporal events e_1 and e_2 in the equation 4.4. The overlap represents the duration d ($d = min(e_1^e, e_2^e) - max(e_1^e, e_2^e)$) where two events e_1 and e_2 appear in the same time windows (in Figure 4.10, e_1 and e_2 overlap between 2sec and 4sec, so $d = 2$). To normalize the overlap between zero and one, we divide d by the time interval td ($td = max(e_1^e, e_2^e) - min(e_1^s, e_2^s)$) corresponding to the union of e_1 and e_2 (in Figure 4.10, $td = 5$).

$$\text{Overlap } (e_1, e_2) = \begin{cases} 0, & \text{if } d < 0 \\ \frac{d}{td}, & \text{otherwise} \end{cases} \tag{4.4}$$

The overlap between a pattern p and a sequence S is the sum of the overlap between the events of p and of S. To normalize (between 0 and 1), we divide this sum by the minimum length between p and S.

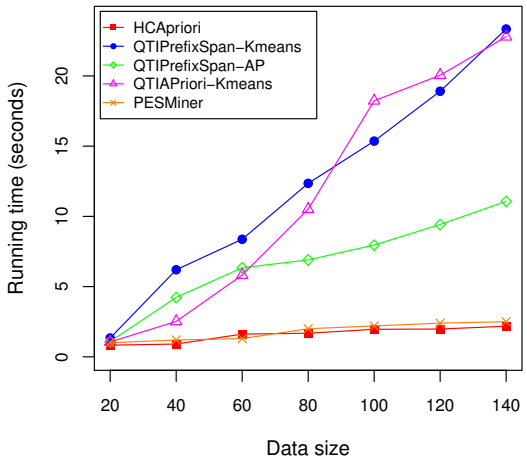


Figure 4.9 – Running time evolution with respect to the dataset size.

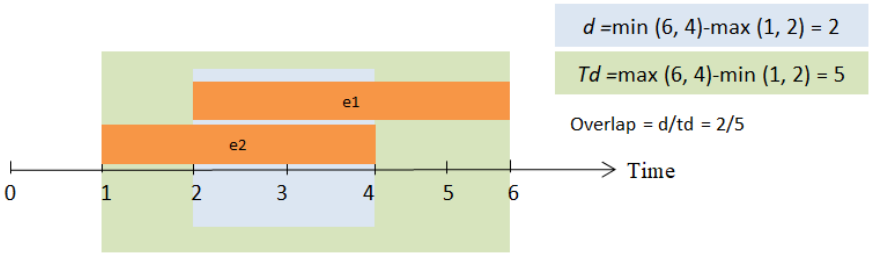


Figure 4.10 – Representation of overlap between two temporal events.

The two new measures $SupOverlap$ and $ConfOverlap$ indicating the support and confidence overlap between a pattern p and a dataset D are given in equation 4.5 and 4.6 respectively.

$$\text{SupportOverlap (p)} = \frac{\sum_{S \in D} \text{overlap (p,S)}}{|D|} \tag{4.5}$$

$$\text{ConfidenceOverlap (p, v)} = \frac{\sum_{S \in D, \text{S expressing v}} \text{overlap (p,S)}}{|\{S \in D : \text{S contains p}\}|} \tag{4.6}$$

4.6 Conclusion

In this chapter, we present an overview of temporal sequence mining. Then, we introduce our algorithm HCApriori that overcomes the limitation of existing algorithms. Because the existing algorithms have been evaluated using synthetic data, they generally fail to efficiently deal with real-word data. First, they do not consider differences of duration of events. In addition, relying on partitioning clustering algorithms (like, Kmeans), distant

events can be merged into a same cluster. Results show that HCApriori allows a better extraction of sequences of non-verbal signals with a significant improvement over the other four state-of-the-art algorithms.

In the next chapter, we conduct a user study to assess if the extracted patterns with our algorithm will be perceived as expressing attitude variations.

Key points:

- Temporal sequence mining algorithms allow extracting exact timing and duration of events.
- We introduce our algorithm HCApriori that overcomes the limitation of existing algorithms by considering differences between event types and increasing cluster homogeneity. HCApriori outperforms the existing algorithms.
- We extend the existing metrics for pattern quality assessment by considering temporality between events.

Chapter 5

Sequence-Based Attitude Variation Modeling

Contents

5.1 Extraction of Relevant Patterns Expressing Attitude Variations **48**
5.1.1 Building Sequence Databases Representing Attitude Variations . 48
5.1.2 Pattern Extraction 54
5.2 Evaluation of the Extracted Patterns **54**
5.2.1 Experimental Design 55
5.2.2 Measures 55
5.2.3 Hypotheses 56
5.2.4 Results 57
5.2.5 Discussion 63
5.3 Conclusion **65**

OUR goal is to develop a virtual agent able to display different attitude variations depending on the interaction context. For example, it should be able to increase its dominance level when interviewing a candidate for a job opening. As presented in Chapter 2.4, interpersonal attitudes are conveyed through non-verbal behaviors (e.g., gaze, facial expression, head movements, etc.). Furthermore, attitudes are not only expressed by specific signals but also by their sequentiality and temporality (occurrence time and duration). Our approach is strongly built on the assumption that temporality is determinant for characterizing attitude variations. To this end, we represent an attitude variation as temporal sequences of non-verbal signals (see Definition 2 in Chapter 4).

Our approach can be summarized as follow: first, we segment a multimodal corpus into four datasets containing sequences related to attitude variations (friendliness increase

or decrease, dominance increase or decrease). Secondly, we apply a temporal sequence mining algorithm to extract, for each attitude variation, the most relevant patterns (subsequences) characterizing this attitude variation. Then, the extracted patterns are simulated within an ECA and evaluated through a perceptive study. Results validate that the extracted patterns express the intended attitude variations. Based on this study, we develop an attitude planner that enables an ECA to communicate with attitude variations.

In this Chapter we describe the different steps we follow to model attitude variation. The attitude planner and its evaluation will be presented in the next Chapter.

5.1 Extraction of Relevant Patterns Expressing Attitude Variations

In this Section, we present our methodology for extracting relevant patterns of non-verbal signals related to attitude variations. First, using a corpus of dyadic interactions, we map the attitude variations to sequences of non-verbal signals. Then, we apply HCApriori to extract the most relevant patterns related to attitude variation.

5.1.1 Building Sequence Databases Representing Attitude Variations

For attitude variation modeling, we use a corpus of job interviews where a recruiter can express different attitudes toward a candidate. Different corpora of job interviews have been collected like TARDIS [Chollet et al., 2014b], HuComTech [Szekrényes, 2014], and the corpus from [Nguyen et al., 2014]. In our work, we use the TARDIS corpus because both interpersonal attitudes and non-verbal behaviors of the recruiter have been annotated. This corpus is composed of three videos showing three couples of candidate-recruiter. The total duration of this corpus is 57 minutes and 32 seconds. This corpus is annotated on two levels: non-verbal behavior and attitude perception of related to the recruiter [Chollet et al., 2014b]. Several modalities of non-verbal behavior have been annotated using the annotation tool Elan [Wittenburg et al., 2006] such as gaze, head directions, head movements, etc. The non-verbal signals that have been annotated for each modality are given in Table 5.1. Using the annotation tool Gtrace developed by Cowie and colleagues [Cowie et al., 2012], the annotation of dominance and friendliness is done continuously. Each annotator annotates only one job interview and one dimension of attitude at a time (dominance or friendliness). The value of annotation ranges from -1 to 1 as indicated in Figure 5.1.

Having these annotations, we segment the non-verbal behaviors based on attitude variations as indicated in Figure 5.3. The attitude variation is defined in Definition 6 and is illustrated in Figure 5.2.

Definition 6 Attitude variation. *An attitude variation v is a tuple $(s, e, vs, ve, value, duration)$, where v^s, resp. v^e, is the starting, resp. the ending, time of the variation. v^{vs} , resp. v^{ve}, is the starting, resp. the ending, attitude value, and v^{value} is the value of variation. Notice that $v^{value} = v^{ve} - v^{vs}$ and $v^{duration} = v^e - v^s$.*

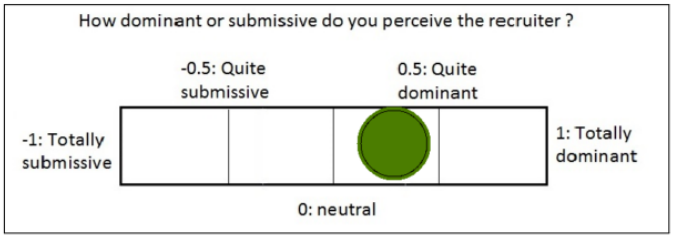


Figure 5.1 – Annotation scale of interpersonal attitude (dominance dimension) [Chollet, 2015].

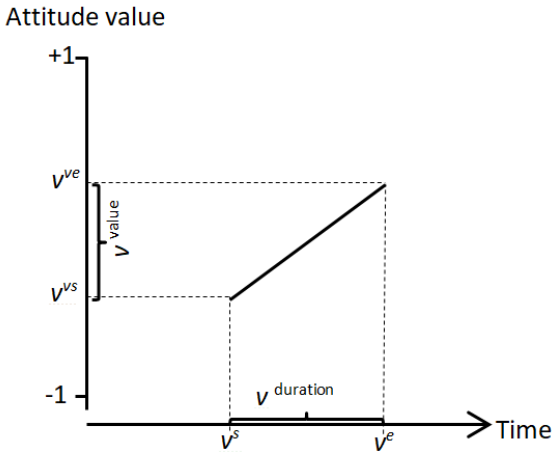


Figure 5.2 – Attitude variation illustration.

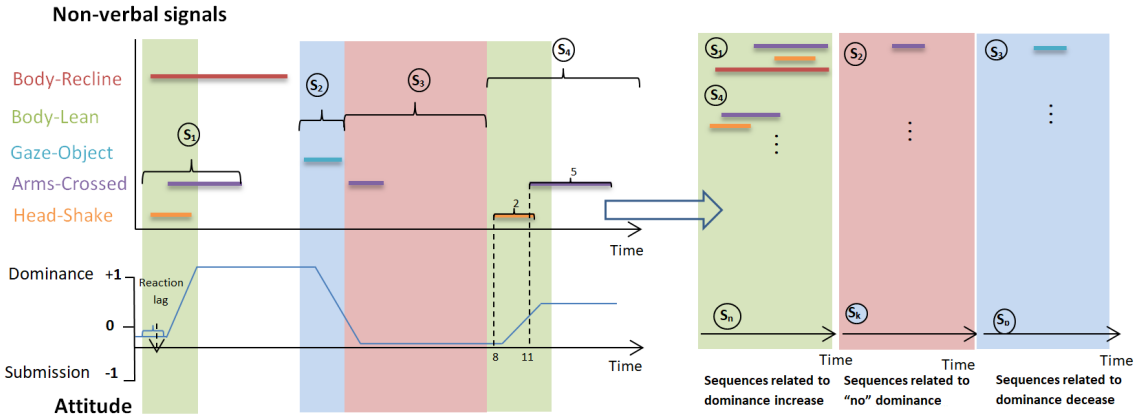


Figure 5.3 – Non-verbal behavior segmentation based on attitude variations.

Modality	Annotated signals
Posture	sitting straight, leaning towards the table, reclining back
Gesture	communicative gestures, object manipulation, adaptor gestures
Hand position	hands on table, hands under table, arms crossed, hands together
Gaze	looking at candidate, looking at object (e.g. table), looking upwards, looking downwards, looking sideways
Head direction	head directed at candidate, head directed upwards, head directed downwards, head directed sideways, head tilted to the side
Head movement	nod, shake
Face	eyebrow raised, eyebrow frowned, smile

Table 5.1 – Annotated non-verbal behaviors in Tardis corpus.

For better annotation quality, we need to consider the reaction lag that specifically occurs in the continuous annotation, as recommended in [Mariooryad and Busso, 2013]. Mariooryad and colleagues demonstrated that the accuracy of emotion recognition improves by more than 7% percent when considering the reaction lag of annotators. Later on, they reported that in the SEMAINE corpus the delay varies from one to six seconds [Mariooryad and Busso, 2015]. These different values of the delay may come from different factors such as the displayed multimodal behaviors, the phenomenon being continuously evaluated or even the annotator's sensitivity. In our study, to choose a value for the reaction lag, we vary its value, lag, from 0 to 6 seconds with a step of 1 sec. For each lag value we compute the accuracy of extracted patterns. The best accuracy results are achieved for $lag = 2$ sec. Thus, we choose this value for the reaction lag.

In table 5.2, we report, for each attitude variation, the mean and standard deviation of v^{vs}, v^{ve}, v^{value} and $v^{duration}$. Figure 5.4 and 5.5 show , respectively, the box plots of the starting value, respectively, the ending value of each attitude variation. Significant differences between the variations can be observed: the variations of friendliness have a larger order of magnitude than the variations of dominance. This can be explained by the fact that the variations of friendliness last generally longer than the variations of dominance (see Figures 5.6 and 5.7).

	v^{vs}		v^{ve}		v^{value}		$v^{duration}$	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
DomInc	0.26	0.22	0.38	0.21	0.12	0.07	5.4	3.5
DomDec	0.34	0.23	0.22	0.22	0.12	0.10	5.9	3.8
FrInc	0.02	0.22	0.22	0.23	0.20	0.15	11.3	8.4
FrDec	0.18	0.22	0.006	0.19	0.17	0.12	12.3	10.9

Table 5.2 – Mean and standard deviations of attitude variation attributes.

For each variation v occurring when the recruiters are speaking, we collect all non-verbal signals that appear during this variation (that is, between $v^s - lag$ and $v^e - lag$). These signals compose a sequence S in which the starting time of each non-verbal signal s is the time difference between the starting time of s and v^s minus lag. For example, in Figure 5.3, the second variation of dominance increase starts at time 9. Then, the

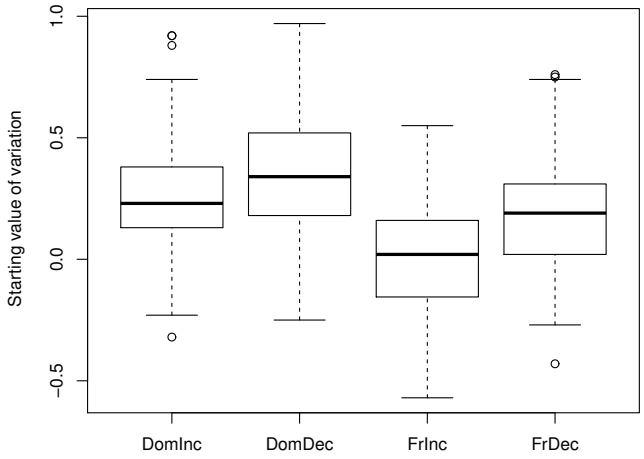


Figure 5.4 – Boxplots of starting value (v^{vs}) for each attitude variation.

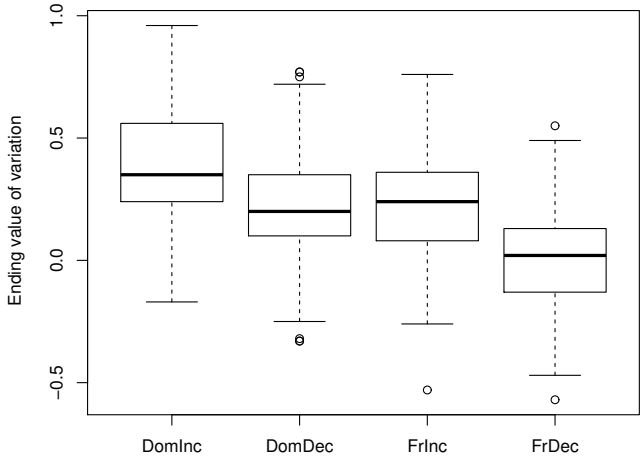


Figure 5.5 – Boxplots of ending value (v^{ve}) for each attitude variation.

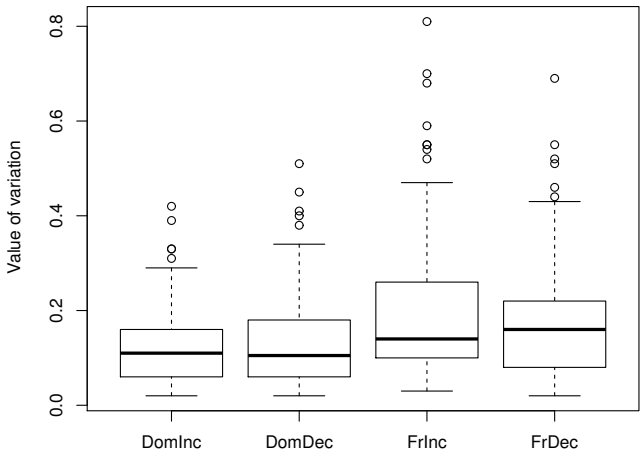


Figure 5.6 – Boxplots of v^{value} for each attitude variation.

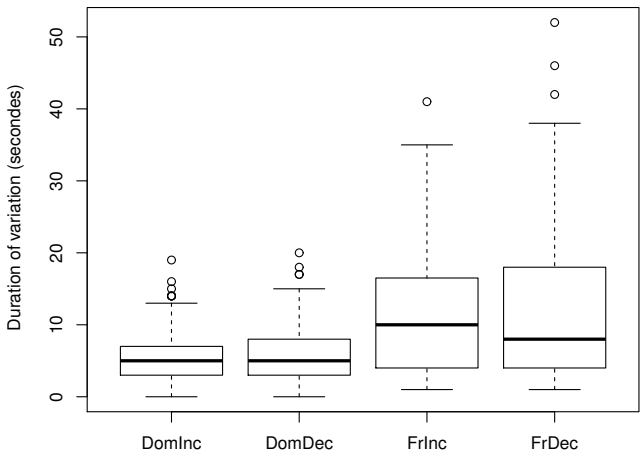


Figure 5.7 – Boxplots of $v^{duration}$ for each attitude variation.

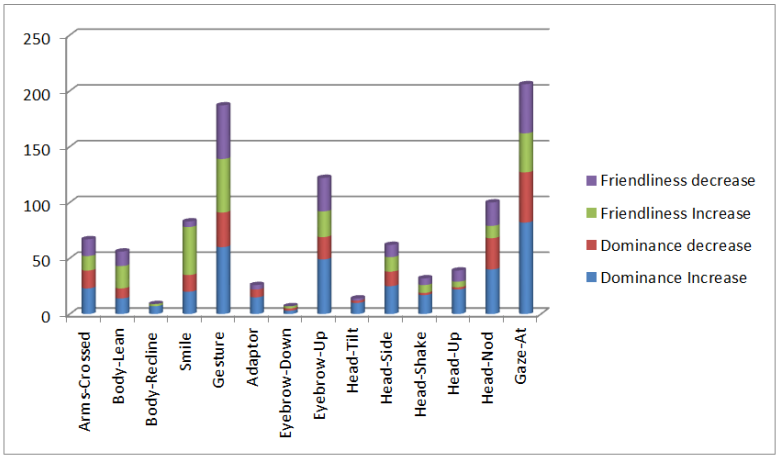


Figure 5.8 – Distribution of non-verbal signals w.r.t. attitude variation.

temporal sequence representing this variation is "*head shake* (1, 2)" followed by "*arms crossed* (4, 5)". This segmentation allows us to build four sets of non-verbal behavior sequences representing four types of attitude variation: dominance increase, dominance decrease, friendliness increase and friendliness decrease. Table 5.3 reports the number of sequences and their average length for each attitude variation when the recruiters are speaking. The sequences occurring during friendliness variations are longer than those occurring during dominance variations. This can be dued as the variations of dominance are shorter than the variations of friendliness (see Figure 5.7).

	Friendliness		Dominance	
	Increase	Decrease	Increase	Decrease
Number of sequences	80	94	143	110
Mean length of sequence (sec.)	10.7	9.4	8	7.8

Table 5.3 – Size of sequences for each attitude variation occurring when the recruiters are speaking.

We also extract sequences of behaviors occurring when "no" attitude is expressed. We define by "no" attitude expression, the segments of the corpus that are marked with an attitude value around zero (for commodity, we take all values between -0.05 and 0.05). We refer to these extracted sequences as "reference" as they express no attitude. Then, since annotators consider one dimension at a time, we obtain 36 and 40 sequences representing respectively "reference dominance" and "reference friendliness".

Figure 5.8 gives the frequencies of nonverbal signals observed during each attitude variation, these results are consistent with the literature (see Chapter 2.4). The following signals are more frequent for dominance increase: *raise head up*, *gesture*, *gaze candidate* and *shake*. Smile occurs primarily during a friendliness increase while adaptor gesture is more present during dominance decrease.

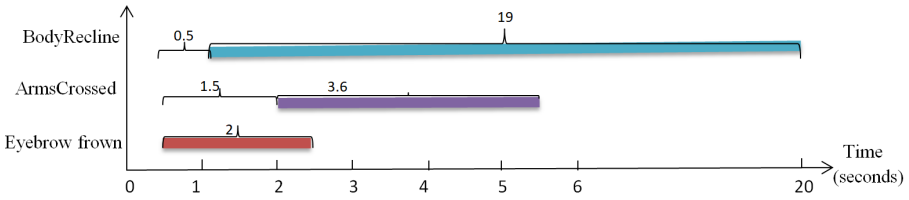


Figure 5.9 – Pattern representing dominance increase.

5.1.2 Pattern Extraction

The previous representation step yields a dataset of temporal sequences. The goal is to extract relevant patterns from each dataset. We perform this extraction based on our algorithm HCApriori. The parameters f_{min} and ϵ are fixed empirically based on prior experiments that gave accuracy value above 70% (see Section 2). This could be reached by fixing f_{min} to 10% of the dataset size and ϵ to 40% of the mean duration of each signal type. The size of extracted patterns for each attitude variation is given in Table 5.4.

FrInc	FrDec	DomInc	DomDec	RefFr	RefDom
210	203	187	156	44	50

Table 5.4 – Size of extracted patterns for each attitude variation as well as the two "reference" attitudes.

Table 5.5 shows an example pattern for each attitude variation as well as the two "reference" attitude. The starting and ending time in seconds of each signal are respectively given between parentheses. The pattern representing dominance increase is illustrated in Figure 5.9, it can be interpreted as follows: 1.5 sec. before recruiter's dominance increases, he frowns his eyebrows for 2 sec. Meanwhile he crosses his arms for 4.6 sec. while leaning backward.

Attitude	Pattern
Dominance increase	EyebrowsFrown (0.5, 2.5), BodyRecline (1, 20), ArmsCrossed (2, 5.6)
Dominance decrease	BodyLean (0, 10.75), GestureAdaptor (6.15, 7.95)
"No" Dominance	EyebrowsRaise (1.6, 2.9), Gesture (2.4, 4.5) HeadsTogether (3.7, 5.6)
Friendliness increase	Smile (1, 3.3), Gesture (2, 4.4), HeadNod (8.1, 10.4)
Friendliness decrease	GaseSide (1.2, 3.2), ArmsCrossed (1, 7.6)
"No" Friendliness	Gesture (0.9, 2.8), EyebrowsUp (3, 5.9), HeadsUnderTable (1.1, 8)

Table 5.5 – Example of patterns obtained with HCApriori.

5.2 Evaluation of the Extracted Patterns

In order to evaluate the non-verbal patterns extracted with our model, we design a perceptive experiment. We hypothesize that an ECA displaying a pattern that represents increase/decrease of an attitude will be evaluated as more/less expressing of this attitude compared to the ECA conveying "no" attitude. First, we describe the experimental pro-

tocol. Then, we analyze the results based on different approaches to assess if attitude variations were correctly recognized.

5.2.1 Experimental Design

We evaluate four different categories of non-verbal patterns denoting four attitude variations: dominance increase (*DomInc*), dominance decrease (*DomDec*), friendliness increase (*FrInc*), and friendliness decrease (*FrDec*). For each of them, we evaluate four non-verbal patterns related to this variation. We also rate two patterns expressing either "no" dominance or "no" friendliness.

Using the virtual agent platform called GRETA-VIB [Pecune et al., 2014], we generate videos showing an agent displaying some patterns, randomly selected from the extracted pattern. We produce a total number of 18 videos: 16 **comparison videos** (4 attitude variations $\times$ 4 patterns) and two **reference videos**: "reference dominance" (denoted *DomRef*) and "reference friendliness" (denoted *FrRef*).

The evaluation follows a two-step process: first participants are asked to view and rate the ECA in the reference video. Then, participants view four pairs of videos where each pair is made of the reference video and a comparison video; they are asked to rate the behavior of the ECA in the comparison video. Participants are randomly assigned to one condition in which the ECA displayed patterns expressing one given attitude variation. Videos appeared automatically once participants view the whole video and answer all questions. The order of videos is shown according to a latin square design to control first-order carryover effects [Bradley, 1958].

5.2.2 Measures

Participants evaluate their perception of agent's attitude along several adjectives. To find the most relevant adjectives that characterize the perception of an attitude, we conduct a detailed literature review of different use cases of the interpersonal circumplex (IPC) measurements. The Interpersonal Check List (ICL) [Leary, 1957] and the Interpersonal Adjective Scales (IAS) [Wiggins, 1979] are two measures for representing interpersonal traits. To measure the perception of attitude, previous researches relied on either IAS [Cafaro et al., 2016a, Janssoone, 2016, Pecune, 2016] or ICL [Op Den Akker et al., 2013]. However, only a limited number of adjectives have been used. For example, Chollet *et al.* used only two variables *friendly* and *dominant* [Chollet et al., 2014b]. In [Cafaro et al., 2016a], three items have been adopted to assess dominance. In our work, we use a combination of both IAS and ICL.

For sake of simplicity, two adjectives with the highest factors in IPC and in IAS, are selected from the analysis done respectively in [Leary, 1957] for ICL and in [Wiggins et al., 1988] for IAS. Thus, we select two adjectives representing high dominance (forceful (ICL), assertive (IAS)), high friendliness (helpful (ICL), tender (IAS)), and so on for the remaining octants (cf. Table 5.6). In total, we use 16 adjectives: 8 adjectives from IAS and 8 from ICL as listed in table 5.6.

	ICL	IAS
PA	forceful	assertive
BC	compete	aggressive
DE	defiant	arrogant
FG	withdrawn	distant
HI	unauthoritative	timid
JK	depend	cooperative
LM	helpful	tender
NO	leader-like	cheerful

Table 5.6 – Selected adjectives from ICL and IAS.

Unlike the previous studies where participants only rated the perception of one attitude dimension at a time [Chollet et al., 2014b, Ravenet et al., 2015, Cafaro et al., 2016a, Janssoone, 2016, Pecune, 2016], we asked the participants to rate the perception of the two dimensions simultaneously to discover the relationship that may exist between the two dimensions such as halo and compensation effect. Compensation effect is a negative relationship between two dimensions of social judgment [Yzerbyt et al., 2008]. A halo effect is a positive relationship between two dimensions, i.e., changing one dimension involves the change of the other dimension in the same direction [Yzerbyt et al., 2008]. For this purpose, participants rated the behavior of the agent by answering 16 questions related to the 16 selected adjectives (cf. table 5.6): *in your opinion, is the behavior of the virtual character assertive?*. All answers are on a 5-point labeled Likert scale as indicated in Figure 5.10 (1 =“strongly disagree”, 2 = “partially disagree”, 3 = “neutral”, 4 = “partially agree”, and 5 = “strongly agree”). In order to detect and filter out the participants who randomly responded to questions, we also ask a trap question about the color of the agent’s hair (the ECA used in this experiment is blond).

5.2.3 Hypotheses

Our hypotheses are:

- **H.Ref**: for *DomRef* and for *FrRef*, the ECA will be evaluated as expressing “no” attitude;
- **H.Dom**: for *DomInc*, the ECA will be evaluated as **more dominant** compared to the ECA in *DomRef*;
- **H.Sub**: for *DomDec*, the ECA will be evaluated as **more submissive** compared to the ECA in *DomRef*;
- **H.Fr**: for *FrInc*, the ECA will be evaluated as **more friendly** compared to the ECA in *FrRef*;
- **H.Hos**: for *FrDec*, the ECA will be perceived as **more hostile** compared to the ECA in *FrRef*.

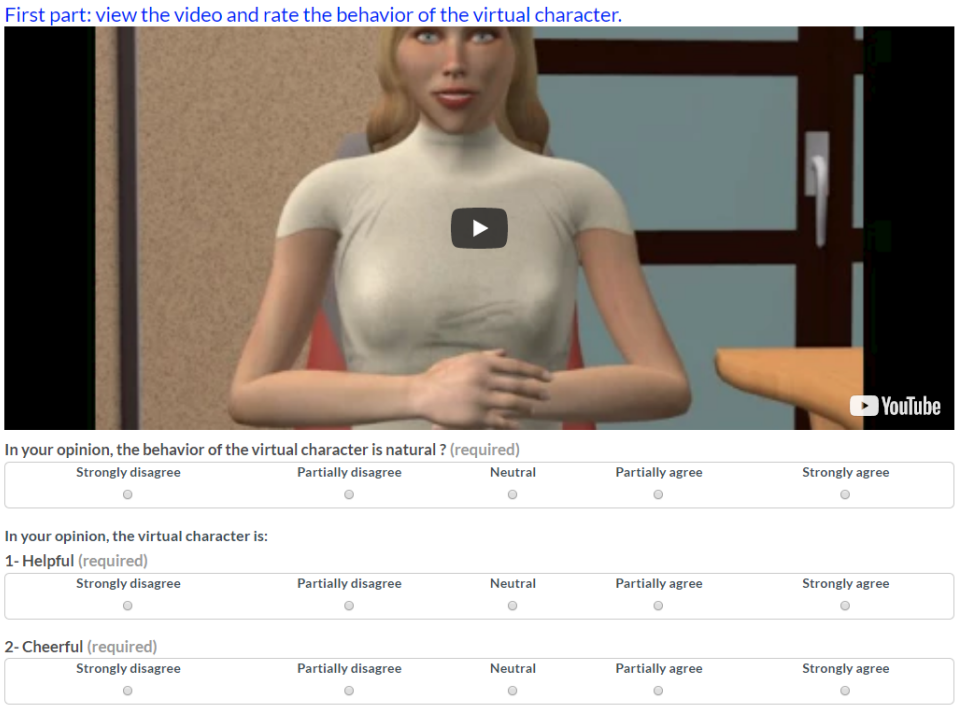


Figure 5.10 – A screenshot from CrowdFlower evaluation platform.

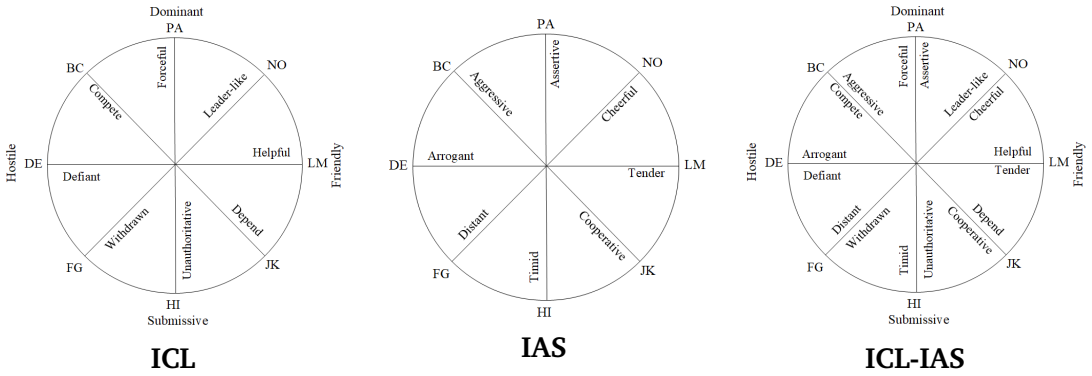


Figure 5.11 – Interpersonal adjectives from ICL and IAS used in our experiment and their placement in the interpersonal circumplex.

5.2.4 Results

We recruit a total of 64 participants via Crowdflower, 42% of them are between 21 and 30 years old, 85% are male, 53% have a master level and 57% are Spanish. We analyze the results in three different ways by: (1) plotting the results on the IPC, (2) investigating significance of the results, and (3) computing the recognition rate of attitude variations.

5.2.4.1 Circular Profile and Vector Scoring

To plot the results on the IPC, we follow the procedure described in [Locke, 2012] (See Section 2.3.2). It consists on averaging the eight octant scores to obtain the general factor

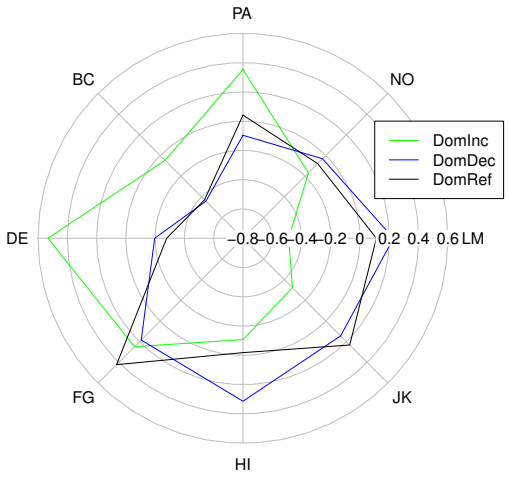


Figure 5.12 – Plotting the ipsatized scores for *DomInc*, *DomDec* and *Ref*.

score. Then, each octant score is ipsatized by subtracting the general factor. Finally, the ipsatized scores are plotted on the IPC.

Figures 5.12 and 5.13 plot the ipsatized scores for each condition. We can observe that:

- For *DomInc* and *FrDec*, the ECA is perceived as: more **dominant (PA)**, more **hostile (DE)** and less **friendly (LM)** compared to the ECA in *Ref*;
- For *DomDec*, the agent is evaluated as more **submissive (HI)** compared to the reference video;
- The ECA's friendliness in *FrInc* is perceived as **equivalent** to ECA's friendliness in reference video.

We also summarize the circular profile of the agent by a vector in the IPC space following the steps described in Section 2.3.2. In our study, the vector angle indicates the predominant attitude of the agent and the vector length shows how intensely the agent expresses this attitude.

Table 5.7 gives the angle and vector length for all conditions. For *Ref* and *FrInc*, the vector angles of the agent are in the LM octant (friendliness region). For *DomInc*, the angle of the ECA is 177.05° and 122.64° for *FrDec*. This means that the ECA in *DomInc* is perceived as more hostile than dominant and conversely the agent in *FrDec* is perceived as more dominant than hostile. For *DomDec*, the agent gets an angle in the submission region (296.51°). Finally, since the vectors length is close to 1 for *DomInc*, *FrInc* and *FrRef*, it means participants perceive the attitude in the agent very clearly [Locke, 2012, Wiggins et al., 1988].

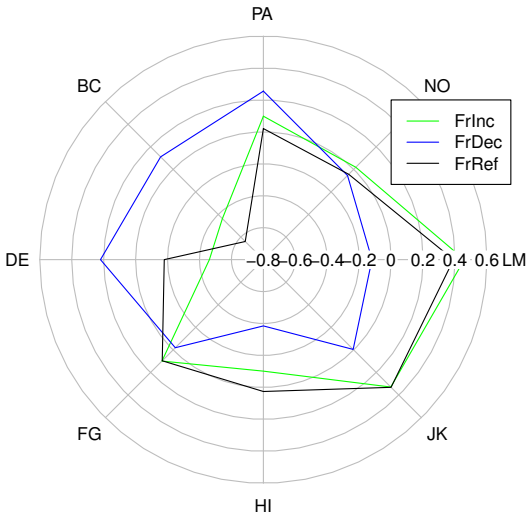


Figure 5.13 – Circular profile of the agent for *FrInc*, *FrDec* and *Ref*.

	FrInc	*FrDec*	*FrRef*	*DomRef*	*DomDec*	*DomInc*
Vector angle	344.2°	122.64°	325.68°	177.05°	296.51°	316.58°
Vector length	0.93	0.74	0.92	0.96	0.82	0.83

Table 5.7 – Angle and vector length of the agent for all conditions.

5.2.4.2 Result Significance

By plotting the agent profile on the IPC, we can visually interpret how the agent is perceived by participants. We also perform statistical tests to investigate if these results are significant or not. In appendix A, we report, for each condition, the mean, standard deviation of independent variables, and distribution of participants' answers over these variable. To check how participants rate the agent in the reference video and in the comparison videos, we conduct a paired Wilcoxon test. The revealed differences are summarized as follows:

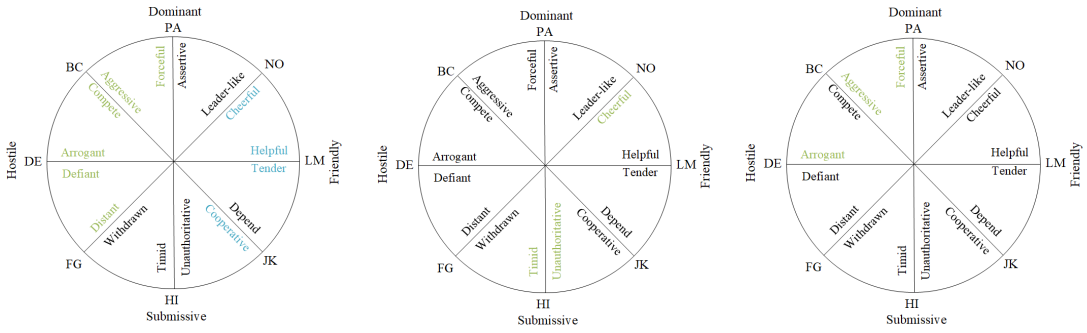
1. ECA in *DomInc* is evaluated as **more dominant** (*aggressive* ($p = .002$) and *forceful* ($p = .01$)) compared to the agent in *Ref*, therefore **H.Dom** is supported. The agent is also perceived as **more hostile** (*compete* ($p = .006$), *arrogant* ($p = .002$), *defiant* ($p = .002$), and *distant* ($p = .006$)) and **less friendly** (*cheerful* ($p = .01$), *helpful* ($p = .002$), *cooperative* ($p = .001$), and *tender* ($p = .005$)) compared to the reference video. We observe that increasing dominance influences not only the perception of dominance but also the perception of friendliness. These results highlight a compensation effect between the perception of dominance and of friendliness: increasing dominance leads to a perception of friendliness decrease and hostility decrease.

2. For *DomDec*, the ECA is evaluated as **more submissive** (*timid* ($p = .01$) and *unauthorative* ($p = .04$)) compared to the agent in *Ref*, thus the hypothesis **H.Sub** is

accepted. The agent is also perceived as **more friendly** (*cheerful* ($p = .01$)) compared to the reference video. These results underline another compensation effect between the two attitude dimensions: decreasing dominance leads to friendliness increase.

3. For *FrDec*, the ECA is perceived as **more hostile** (*arrogant* ($p = .01s$)) compared to the agent in *Ref*, therefore **H.Hos** is validated. In addition, the agent is evaluated as **more dominant** (*aggressive* ($p = .001$) and *forceful* ($p = .001$)) compared to the reference video. Then, we find another compensation effect: decreasing friendliness leads to dominance increase.

4. For *FrInc*, participants rate the ECA's friendliness as **equivalent** to the ECA's friendliness in *Ref*, therefore **H.FR** is rejected.



Dominance increase (a) **Dominance decrease (b)** **Friendliness decrease (c)**

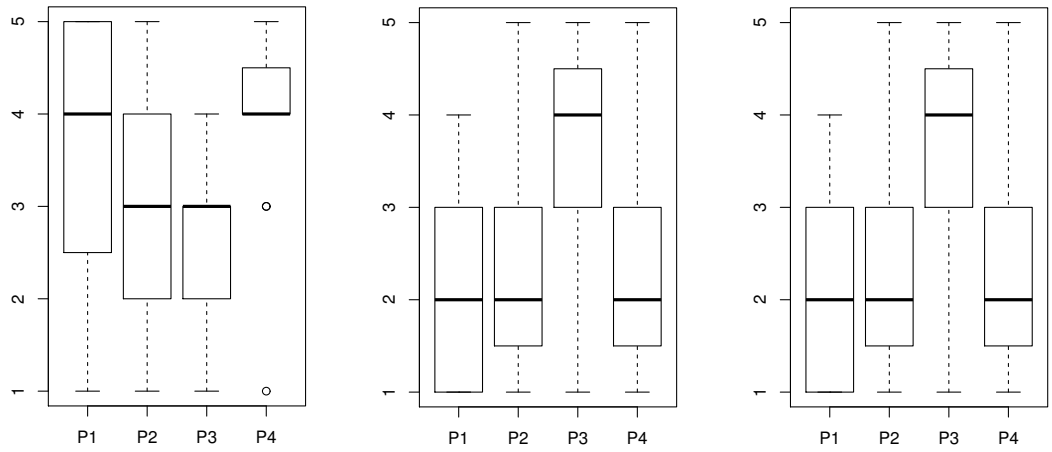
Figure 5.14 – Differences in the ECA perception between the comparison video and the reference video. The blue color indicates that the variable is evaluated as being more expressed in the reference video; the green indicates that the variable is evaluated as being more expressed in the comparison video; and the black denotes no difference between the reference video and the comparison video.

5.2.4.3 Comparison of Patterns Within the Conditions

Pattern	Non-verbal signals
P_1	Body Recline (1, 20), Eyebrow down (2, 4), Arms Crossed (10, 20), Beat (9.25,10.45)
P_2	Beat (0.65, 2.65), Head shake (2.15, 4.26) Beat (5.5, 7.5), Eyebrow up (6.1,8.2)
P_3	Beat (1.1, 3.2), Beat (5, 6.8), Arms crossed (7, 20), Eyebrow down (9.6, 10.9)
P_4	Arms crossed (0.65, 20.85), Beat (0.65, 2.8), Head shake (2.15, 4.15), Head side (5.45, 9.4)

Table 5.8 – The four evaluated patterns for *DomInc*.

We want to understand if a given pattern, from the different patterns we use to evaluate the attitude variations, has an impact in the perception of an attitude change. So for each attitude variation, we evaluate four patterns (p_1, p_2, p_3, p_4). To explore the effects of the four patterns within their respective four attitude variations, we conduct a Friedman test (non-parametric test alternative to the one-way ANOVA with repeated measures). No



(a) Compete **(b) Timid** **(c) Unauthorative**

Figure 5.15 – Boxplots of some variables of the four patterns expressing dominance increase . (1) "strongly disagree", (2) "partially disagree", (3) "neutral", (4) "partially agree", (5) "strongly agree".

significant differences between the four patterns have been detected for friendliness increase and friendliness decrease.

For dominance increase, results revealed a significant difference between the four patterns that characterized this attitude change. This difference concerns the evaluation of three variables: *compete* ($p = 0.003$), *timid* ($p = 0.0003$), and *unauthorative* ($p = 0.002$). Figure 5.15 presents boxplots of these three variables and table 5.8 shows the four patterns representing dominance increase used in the experiment. To understand this difference in perception, we conduct further analysis. Results of the two tests (Friedman and Bonferroni post-hoc) are reported in table 5.9. Bonferroni post-hoc test indicates the pairs of patterns that have been perceived differently and their $p - value$. We observe that the ECA displaying pattern P_3 has been evaluated significantly more *timid* and *unauthorative* than the ECA displaying P_2 and P_4 (see Figure 5.15). We also observe that P_4 has been evaluated as the most expressive for dominance (in term of *compete*) compared to the three other patterns. These differences could be caused by any parameters defining the sequences of behaviors such as the order of signals in the pattern, the signals type, their starting time, or even their duration. For P_4, it could be the presence of the non-verbal signal *head side* or *arms crossing* at the beginning of speech. This hypothesis needs further investigation. To check if the pattern P_4 contributes the most in the recognition of dominance increase, we compare (through a Wilcoxon test) *DomRef* with P_1, P_2, and P_3. The same result has been obtained as when considering all patterns (see Section 5.2.4.3). Thus, all four patterns convey a dominance increase.

Variable	Chi-squared	Df	P	Bonferroni test
Compete	13.9	3	0.003	(P_3, P_4, $p = 0.03$)
Timid	18.21	3	0.0003	(P_1, P_3, $P = 0.02$), (P_3, P_4, $p = 0.03$)
Unauthoritative	14.78	3	0.002	(p_3, p_2, $P = 0.04$), (P_3, p_4, $p = 0.02$)

Table 5.9 – Friedman test and Bonferroni post-hoc test exploring the effects of the different patterns on *DomInc*.

5.2.4.4 Recognition Accuracy of Attitude Variations

In order to assess how accurate is the recognition of the attitude variations, we cast the problem as multi-label classification task, where the predicted class (label) can be one or more among the four attitude variations (*DomInc*, *DomDec*, *FrInc*, and *FrDec*). Thus, we use classical measures from Information Retrieval: recall, precision and F-measure. The recall of a given variable V represents the number of videos expressing V and evaluated as expressing V relative to the total number of videos expressing V (64 videos resulting from 16 participants $\times$ 4 videos). The precision of a given variable V is defined as the number of videos expressing V and evaluated as expressing V relative to the total number of videos evaluated as expressing V. We consider that a given video is evaluated as expressing V if the participant's response for V is either "partially agree" or "totally agree". For example, 31 videos representing dominance increase are assigned to *"forceful"* then the recall of *"forceful"* is 48.43% (31/64). Also, 6 videos representing dominance decrease, 7 videos representing friendliness increase, and 48 representing friendliness decrease are rated as expressing the variable *"forceful"*. Consequently, its precision is 33.69% (31/(31+6+7+48)). F-measure is finally computed as the harmonic mean of recall and precision.

Each measure is calculated for each attitude variation (condition) by averaging the results obtained from its representative variables (given between parenthesis in Table 5.10).

	DomInc (*PA*)	DomDec (*HI*)	FrInc (*LM*)	FrDec (*DE*)
Recall	39%	40%	35%	35%
Precision	34%	43%	36%	31%
F-measure	36%	41%	35%	33%

Table 5.10 – Recall, precision, and F-measure for each attitude variations.

As we can see on Table 5.10, the best results are achieved for *DomDec*. The recall for the four conditions is less than 50% which means that only less than half of videos expressing a given attitude variation are recognized by participants as expressing this attitude variation. The precision is less than 50% for all attitude variations, which means that, for each attitude variation, more than half of the videos annotated as expressing this variation are actually assigned to another attitude variation. In Table 5.11, we report the distribution of the predictions over the actual conditions of the predictions. A cell in this Table (where actual=A and predicted $=B$) gives the number of videos actually expressing A and evaluated as expressing B. From these results, we validate the compensation

effects given in Section 5.2.4.3. In addition, we observe that for both reference videos, participants perceive the agent to be **friendly** but not **hostile**, nor **submissive**.

		Predicted			
		DomInc	*DomDec*	*FrInc*	*FrDec*
Actual	*DomInc*	**39%**	22%	20%	**36%**
	DomDec	25%	**40%**	**40%**	15%
	FrInc	13%	22%	**35%**	26%
	FrDec	**51%**	18%	7%	**35%**
	DomRef	34%	18%	**52%**	4%
	FrRef	20%	12%	**43%**	10%

Table 5.11 – Distribution of the predictions over the actual conditions. The predictions highlighting the compensation effects given in Section 5.2.4.3 and the friendliness perception of the agent in *DomRef* and *FrRef*.

5.2.5 Discussion

The reference videos are generated from the non-verbal sequences that were perceived with attitude values close to zero. We assume that the agent in these videos would be perceived as expressing "no" attitude. To our surprise, the result of the study shows that the agent is evaluated as friendly which invalidates the hypothesis **H.Ref**. We find no significant differences in the perception of the agent in the reference videos and in the *FrInc* condition. The hypothesis **H.Fr** is not validated. An explanation could be that, since the agent in the reference videos is already evaluated as friendly, the agent in the *FrInc* is not perceived as being significantly more friendly than in the comparison videos (*FrRef*). The three other hypotheses, **H.Dom**, **H.Sub** and **H.Hos**, are validated as there are significant differences in the perception of the agent in the *DomInc*, the *DomDec* and in the *FrDec* conditions.

For *DomInc*, we find a main effect of the four evaluated patterns on the perception of the ECA's attitude. The revealed differences concern 3 out of the 16 adjectives. These differences can be caused by any parameters defining the pattern of behaviors. Further study needs to be conducted to understand this. However, we look at understanding the impact on each stimuli the perception of the attitude variation. Thus we check if the difference in perception come from the perception of the agent in one of the four videos used as stimuli; that is, if a particular pattern of behavior can cause these differences in perception. To test this, we redo the statistical test four times on 4 videos, each time ignoring one of the four videos. In each case, we do not find any changes in the perception of the agent. We conclude that all four patterns are perceived as conveying a dominance increase.

According to the representation of the attitudes on the interpersonal circumplex, the two poles of an attitude dimension (dominance/submission, Friendliness/hostility) are symmetrical with respect to the center of the circumplex. As a result, it is expected that the increase of an attitude toward a given pole would result in a decrease in the perception of

the opposite pole. For example, an increase of friendliness would decrease the perception of hostility and vice versa. Based on the circular profile (see Figures 5.12 and 5.13), for both poles of each attitude dimension, this relationship is observed in both directions of the attitude variations. However, it is not statistically significant.

Several works on attitude modeling rely on the assumption that there is a compensation effect between the two attitude dimensions. To compute which social attitudes an agent conveys to its interlocutor [Kasap et al., 2009, Pecune et al., 2016] defined rules such as positive emotions felt and conveyed by the agent increase its friendliness and decrease its dominance toward the user. Vice versa, negative emotions decrease its friendliness and increase its dominance [Kasap et al., 2009, Pecune et al., 2016]. Others works rely on the interpersonal complementary to model the attitude of agents [Ravenet et al., 2015]. According to this theory, two persons should express complementary or anti-complementary attitudes in order to maintain an interaction: expressing similar attitudes on the friendliness dimension and opposite attitudes on the dominance dimension [Leary, 1957, Kiesler, 1996]. But, to the best of our knowledge, there is no studies, in term of perception, on the interrelation of the interpersonal attitudes dimensions. To study this interrelation, we evaluate both dimensions of attitudes at the same time. Doing so allows us to underline a compensation effect between the perception of dominance and of friendliness drawn from the following observations:

- dominance increase leads to a perception of friendliness decrease;
- dominance decrease leads to a perception of friendliness increase;
- friendliness decrease leads to a perception of dominance increase.

In table 6.4 we summarize the relationships revealed in our study between attitude variations and attitude perception. For each relationship, we indicate if it is statistically significant (Stat.), if it is validated based on the circular profile of the agent (IPC); we also report the direction (Direct.) of the relationship (increase or decrease). For instance, increasing dominance raises the perception of dominance and hostility and reduces the perception of friendliness and submission. However, the underlining interrelation between dominance and submission is not statistically significant.

	Dominance			Submission			Friendliness			Hostility		
	IPC	Stat.	Direct.	IPC	Stat.	Direct.	IPC	Stat.	Direct.	IPC	Stat.	Direct.
DomInc	✓	✓	↗	✓	x	↘	✓	✓	↘	✓	✓	↗
DomDec	✓	x	↘	✓	✓	↗	✓	✓	↗	-	-	-
FrInc	-	-	-	-	-	-	✓	x	↗	✓	x	↘
FrDec	✓	✓	↗	✓	x	↘	✓	x	↘	✓	✓	↗

Table 5.12 – Relationships between attitude variations and attitude perception.

We observe that there is high correlation between the perception of dominance increase (*DomInc*) and the perception of friendliness decrease (*FrDec*). An explanation is that some non-verbal signals have the same effect on the perception of dominance and of hostility [Knutson, 1996, Tiedens et al., 2000, Carney et al., 2005, Ravenet et al.,

2013]. For example, both dominance and hostility are characterized by a negative facial expression and no gaze avoidance [Knutson, 1996, Tiedens et al., 2000, Carney et al., 2005, Ravenet et al., 2013].

3 out of our 5 hypotheses (**H.Dom**, **H.Sub** and **H.Hos**) have been validated. So the sequences expressing the corresponding attitude variations are properly recognized. This supports our assumption that attitude variations can be represented as sequences of temporally ordered non-verbal signals. Our next step is to use the extracted sequences to build an attitude planner for virtual agents.

5.3 Conclusion

In this Chapter, we present our methodology to model attitude variations as sequences of non-verbal signals. First, we build a database of sequences representing different attitude variations. Then, we apply HCApriori to extract frequent patterns related to each attitude variation. We also conduct a user study to evaluate the perception of the extracted patters. Results have been plotted on the ICP and analyzed with statistical tests. Furthermore, we compute recall, precision and F-measure to better understand the qualitative results. Using these three methods to analyze the results allows us to better understand them through qualitative and quantitative measures.

Take home

- We presented our methodology to extract relevant patterns related to different attitude variations.
- Extracted patterns are evaluated through a perceptive study.
- Results are plotted on IPC, analyzed based on statistical tests and Information Retrieval metrics.
- All extracted patterns (attitude variation) are correctly perceived.
- To our surprise, the patterns representing “no" attitude are perceived as conveying friendliness.
- We found a compensation effect between the two attitude dimensions.
- We found high correlation between dominance increase and friendliness decrease.

Chapter 6

Attitude Planner

Contents

6.1 GRETA-VIB **68**

6.2 Sequential Attitude Planner Model **70**

6.2.1 Intention Sequence Generation 70

6.2.2 Attitude Sequence Selection 71

6.2.3 Intention Sequence Enrichment 72

6.2.4 Signal Replacement 73

6.3 Evaluation **73**

6.3.1 Experimental Design 74

6.3.2 Results 74

6.4 Discussion **77**

6.5 Conclusion **79**

GRETA-VIB, platform has been developed to support the creation of socio-emotional ECAs [Pecune et al., 2014]. In this platform the agent displays an utterances augmented with communicative functions and emotional states. The ECA's attitude plays a key role in order to successful the interaction goal [Kasap et al., 2009, Ochs et al., 2010, Pecune et al., 2016]. In this chapter, we describe how we enhanced GRETA-VIB platform with our attitude model described in Chapter 5. To this end, we develop an attitude planner that combines the attitude variation of the ECA along with its communicative intentions and emotions. We first present GRETA-VIB platform. Then, we describe the process behind our attitude planner. Finally, we evaluate the attitude planner and we discuss the results.

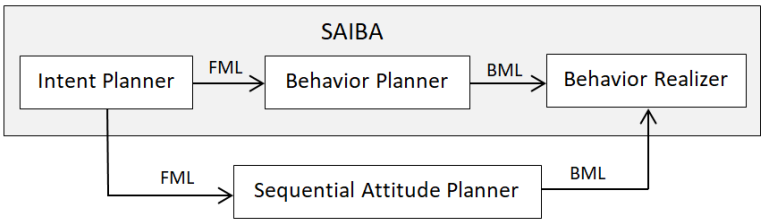


Figure 6.1 – SAIBA architecture enhanced with the new integrated module: *Sequential Attitude Planner*.

```
<?xml version="1.0" encoding="UTF-8"?>
<fml-apml composition="append" social_attitude="DomInc">
    <bml>
        <speech id="s1" language="french" start="0.0"
                voice="openmary">
            <tm id="tm0"/>
            To begin,
            <tm id="tm1"/>
            can you
            <tm id="tm2"/>
            present yourself
            <tm id="tm3"/>
            please  ?
            <tm id="tm4"/>

            <pitchaccent id="pa1" start="s1:tm1" end="s1:tm2" type="Hstar"/>

            <boundary id="b2" start="s1:tm4" end="s1:tm4+0.5" type="HH"/>
        </speech>
    </bml>
    <fml>
        <performative id="d1" importance="1.0" start="s1:tm1" end="s1:tm3" type="ask"/>
    </fml>
</fml-apml>
```

Figure 6.2 – Example of a FML file.

6.1 GRETA-VIB

GRETA-VIB is built based on the SAIBA framework [Vilhjalmsson et al., 2007] whose architecture is illustrated in figure 6.1. First, the "intent planner" generates the communicative intentions of the ECA (what the agent intends to communicate such as its speech and emotion). Communicative intentions are represented in the Functional Markup Language (FML) [Heylen et al., 2008]. Figure 6.2 gives an example of an FML file that contains the sentence: "*To begin, can you present yourself please?*". In this example, we have three communicative intentions: two that are linked to prosody, emphasis (*pitchaccent*) and question marker (*Boundary tone*), and ask a question (*performative*). For each intention we indicate its starting and ending time.

Secondly, the "behavior planner" translates these communicative intentions into a set of multimodal signals (e.g., gesture, facial expressions). This planner is based on the framework described in [Mancini and Pelachaud, 2007] and instantiates an intention into a set of non-verbal behaviors called "behavior set". For example, the communicative intention *greet* can be expressed by either a *hand gesture*, a *facial expression (smile)*, or *raise eyebrows*. Figure 6.3 represents the "behavior" set to indicate the behaviors that convey

```
<behaviorset name="boundary-HH">
    <signals>
        <signal id="1" name="faceexp" modality="face">
            <alternative name="faceexp=neutral" probability="0.6"/>
            <alternative name="faceexp=raise_eyebrows" probability="0.3"/>
        </signal>
        <signal id="2" name="gaze=look_at" modality="gaze"/>
    </signals>
</behaviorset>
```

Figure 6.3 – “Behavior set” for question mark *boundary-HH*.

```
<?xml version="1.0" encoding="UTF-8"?>
<bml xmlns="http://www.mindmakers.org/projects/BML" character="Greta"
    composition="append" id="bml1" reaction_duration="NONE" reaction_type="NONE">
    <speech xmlns="" id="s1" language="french" start="0.0" voice="openmary">
        <tm id="tm0"/>
            To begin,
            <tm id="tm1"/>
            can you
            <tm
            id="tm2"/>
            present yourself
            <tm
            id="tm3"/>
            please  ?
            <tm id="tm4"/>
        <pitchaccent end="s1:tm2" id="pa1" start="s1:tm1" type="Hstar"/>
        <boundary end="s1:tm4+0.5" id="b2" start="s1:tm4" type="HH"/>
    </speech>
    <head end="1.249" id="pa1_0" lexeme="Down_Aside_Right" start="0.809">
    <face amount="1.000" end="1.249" id="pa1_1" start="0.809">
    <gesture id="d1_0" lexeme="vouSoumia" ready="0.809" relax="2.253">
    <head end="3.383" id="b2_0" lexeme="Up_Right" start="2.883">
</bml>
```

Figure 6.4 – Example of a BML file.

a question mark *boundary-HH*. As we can see, this function can be translated using two combinations of gaze and face: *GazeAt* and *NeutralFace*, or *GazeAt* and *EyebrowRaise*. The planner selects one combination considering the behavior preferences of the agent and the compatibility with surrounding communicative intentions [Mancini and Pelachaud, 2007]. The selected behaviors are represented in the Behavior Markup Language (BML) format [Vilhjalmsson et al., 2007]. Figure 6.4 shows the BML file corresponding to the FML file given in Figure 6.2.

Finally, the BML tags are transformed into the final animation of the ECA by the “behavior realizer”. The “behavior realizer” is connected to several modules for computing the different animations of the agent’s face, lips, body, and audio (cf. Figure 6.5).

To model social attitude within the agent platform, we replace the “behavior planner" with a new one called “Sequential Attitude Planner". It takes as input an FML file (containing the utterance to be said by the agent) as well as the intentions and the attitude variation that the ECA should express toward the user. “Sequential Attitude Planner" generates the agent’s behavior according to its intentions and attitudes. We update FML format by adding a new tag that represents the attitude variation of the agent (see Figure 6.2).

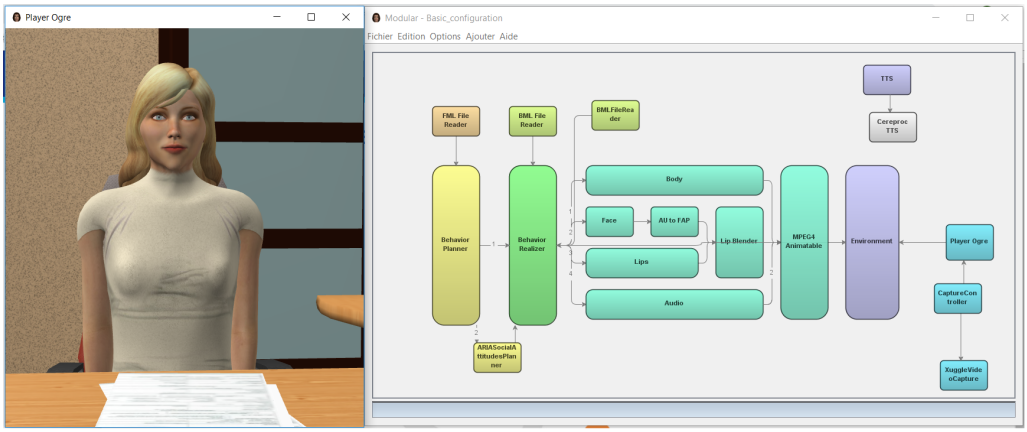


Figure 6.5 – GRETA-VIB platform.

The “Sequential Attitude Planner" is a four step process ranging from sequence generation, selection and enrichment, to the signal replacement. The following Section goes through each of these steps.

6.2 Sequential Attitude Planner Model

To enrich the communication skills of an ECA, we need to combine the sequences of non-verbal signals expressing an attitude variation with those communicating its other communicative intentions. For that, the “Sequential Attitude Planner” follows four steps as indicated in figure 6.6. First, it translates the communicative intentions into a sequence of non-verbal behaviors called *intention sequence*. Then, it selects, from the extracted sequences linked to attitude variations, the most relevant one (*attitude sequence*) representing the given attitude variation of the agent. Then, the *intention sequence* is enhanced with new signals from the *attitude sequence*. Finally, the attitude planner merges the signals from both sequences, the *intention sequence* and the *attitude sequence*, in order to obtain the final signals to be displayed by the agent. Figure 6.7 illustrates the four steps of the “Sequential Attitude Planner".

6.2.1 Intention Sequence Generation

The communicative intent planner generates a sequence of non-verbal behaviors expressing the communicative intentions specified in the input FML file. Once all communicative intentions are instantiated, we obtain a sequence of multimodal behaviors that we call *intention sequence* (S_{int}). In the example described in Figure 6.7.1, the FML contains three intentions: *Emphasis*, *Performative (ask question)*, and *Question marker (Boundary tone)* associated to the sequence S_{int}: *head*, *gesture*, and *face* and *gaze*.

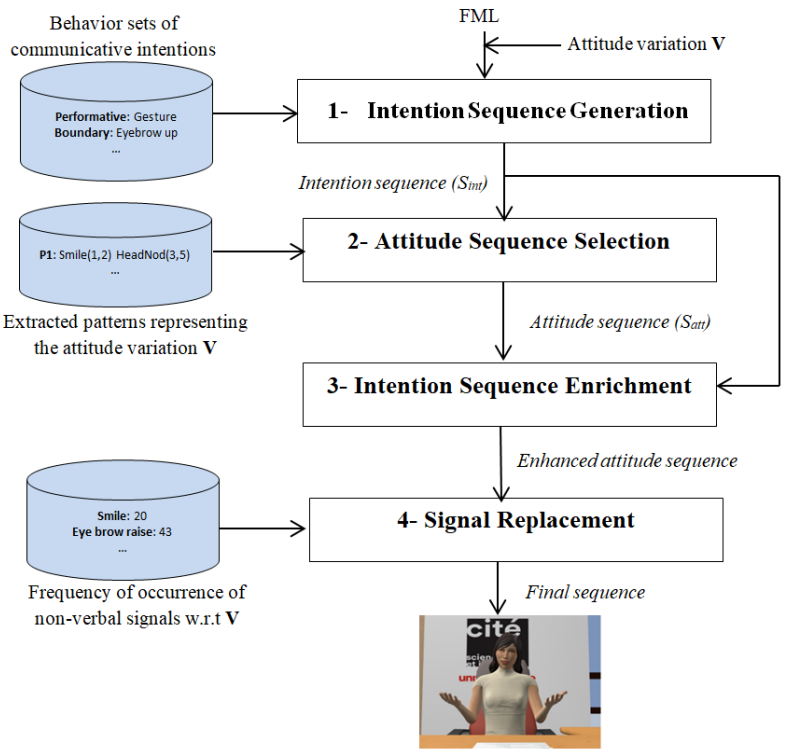


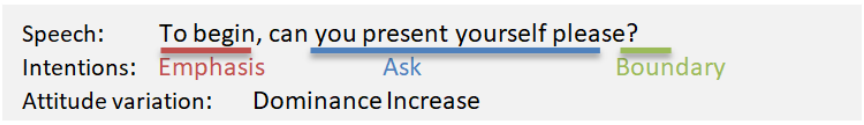
Figure 6.6 – Outline of the sequential attitude planning model.

6.2.2 Attitude Sequence Selection

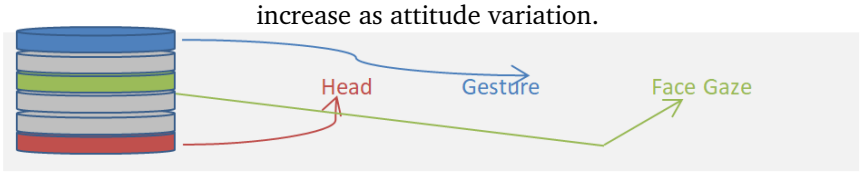
Once the communicative intentions are instantiated, the next step is to choose, from the extracted patterns (cf. Section 5.1.2), the most appropriate sequence *attitude sequence* (S_{att}) conveying the desired attitude variation (V). Appropriateness of S_{att} is defined here as the most representative sequence for conveying the attitude variation V and as the most similar to S_{int}. The representativity of S_{att} for expressing the attitude variation V is evaluated in term of support (Eq. 4.2 and Eq. 4.5) and confidence (Eq. 4.3 and Eq. 4.6) as indicated in Eq. 6.1.

$$\begin{aligned} AttitudeRep(S_{att}, V) = Conf(S_{att}, V) \times Sup(S_{att}) \\ \times ConfOverlap(S_{att}, V) \times SupOverlap(S_{att}) \end{aligned} \tag{6.1}$$

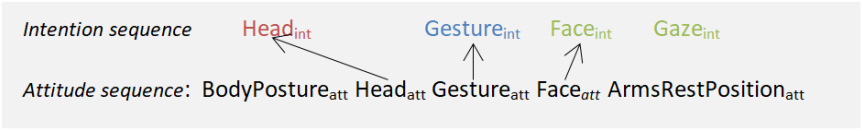
The similarity between S_{int} and S_{att} is evaluated in terms of the presence of multimodal behaviors and of their temporality ($overlap$) as defined in Equation 6.2. $SimType$ returns the number of behaviors from sequences S_{int} and S_{att} that are of the same modality and $Overlap(S_{int}, S_{att})$ represents the time overlap between S_{int} and S_{att}. We consider several non-verbal modalities: *gesture*, *gaze*, *head*, *postures*. Within each modality, we have several signal types. For example, for head modality we have: *head up*, *head down*, *head tilt* and *head side*, etc. In the example (Figure 6.7.2), $SimilarityType = 3$, we have



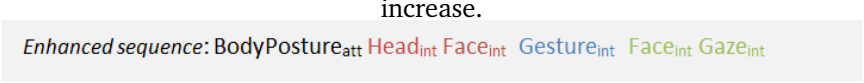
Example of an input FML message containing 3 communicative intentions and dominance increase as attitude variation.



(1) Generation of intention sequence, using the behavior sets of the input communicative intentions.



(2) Attitude sequence selection from the extracted frequent patterns representing dominance increase.



(3) Intention sequence enrichment.

(4) Signal replacement.

Figure 6.7 – An illustrative example for the sequential attitude planning model.

three non-verbal modalities (*head*, *gesture*, and *facial expression*) that are present in both S_{int} and S_{att}.

$$Similarity(S_{int}, S_{att}) = SimType(S_{int}, S_{att}) \times Overlap(S_{int}, S_{att}) \tag{6.2}$$

In this step, we also associate behavior from *intention sequence* to their mapping behaviors in *attitude sequence*: $head_{int} \rightarrow head_{att}$, $gesture_{int} \rightarrow gesture_{att}$, and $face_{int} \rightarrow face_{att}$.

6.2.3 Intention Sequence Enrichment

In the next step, we enrich the set of communicative behaviors with the set of attitude behaviors. It is obtained by merging both S_{int} and S_{att}: each behavior b_{att} in the S_{att} that does not appear in the S_{int} is added to the S_{int}. Using the same example (Figure 6.7.3), we add the behaviors *body posture* and *arms rest position* to the S_{int}. The timing of b_{att} is still the same if it does not overlap with another signal b_{int} in the *intention sequence* from the same modality of b_{att}. Otherwise, we adjust the timing of b_{att} to allow the agent to display the behavior b_{int} as indicated in Equation 6.3 and illustrated in Figure 6.8 and 6.9. For example, if the agent has the intention to ask a question instantiated by a gesture (b_{int}) from 0 to 1.5 sec. and, in the *attitude sequence*, we have another gesture (b_{att}) from 1 to 4 sec. then, the agent will play the gesture b_{att} when it finishes doing b_{int} (at 1.5 sec. instead of 1 sec.).

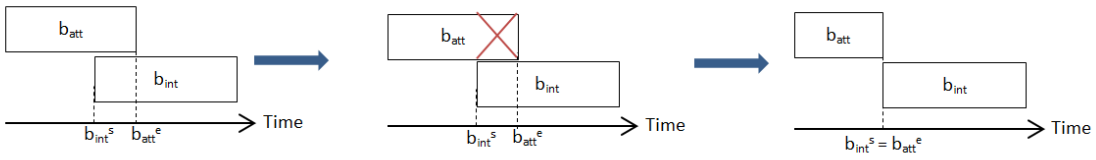


Figure 6.8 – Time adjustment of b_{att} when $b^e_{att} > b^s_{int}$

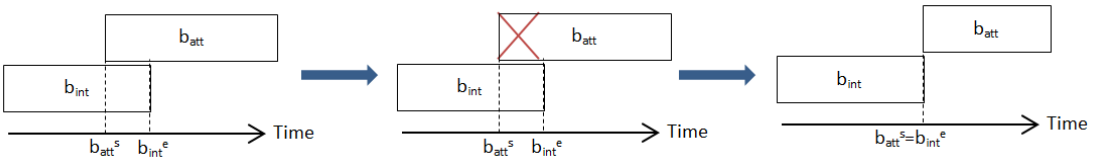


Figure 6.9 – Time adjustment of b_{att} when $b^e_{att} < b^s_{int}$

$$b^e_{att} = \begin{cases} b^s_{int} & \text{if overlap } (b_{att}, b_{int}) > 0 \text{ and } b^e_{att} > b^s_{int} \\ b^e_{int} & \text{if overlap } (b_{att}, b_{int}) > 0 \text{ and } b^e_{att} < b^s_{int} \\ b^e_{att}, & \text{otherwise} \end{cases} \tag{6.3}$$

6.2.4 Signal Replacement

In order to represent the relationship between non-verbal behaviors and attitude variations, we compute the frequency of occurrence of a given behavior b with respect to a given attitude variation V. We consider that a behavior b_1 is more representative of an attitude variation V than a behavior b_2 if the frequency of occurrence of b_1 is higher than the frequency of occurrence of b_2.

Finally, our model will replace each behavior b_{int} of the *intention sequence* with its mapped behavior b_{att} in the *attitude sequence* if the frequency of b_{att} is higher than the frequency of b_{int} (Figure 6.7.4). In the example described in Figure 6.7.4, the attitude planner will choose b_{att} (*Eyebrows-Frown*) as final signal because the frequency of this signal for dominance increase is higher than the frequency of $face_{int}$.

6.3 Evaluation

The *Sequential Attitude Planner* generates the behavior of a virtual agent according to its intentions and the attitude variation it should express. In this section, we report on the perceptive study we conducted to evaluate the behaviors of an ECA generated with our *Attitude Planner*. Since we have used a job interview corpus to extract the most relevant non-verbal sequences related to different variations of the recruiter's social attitudes, we keep a similar scenario for this evaluation study. Thus, the ECA plays the role of recruiter interviewing for a job opening.

6.3.1 Experimental Design

We design an empirical experiment in which participants compare a set of video pairs. Each pair is made of a video of the virtual recruiter with no attitude variation and a second video with an attitude variation. We choose seven sentences (questions) that have a rather "neutral" verbal content. An example of sentence is: *if we decided to offer you this job, when would you be ready to start?*. Seven reference videos (*Ref*) are generated without our *sequential attitude planner* (i.e. displaying no attitude change), and 28 others with our *sequential attitude planner* (4 attitude variations $\times$ 7 sentences).

We follow the same structure for the evaluation study as the previous one (see Section 5.2). We evaluate five experimental conditions corresponding to the four attitude variations: dominance increase (*DomInc*), dominance decrease (*DomDec*), friendliness increase (*FrInc*), friendliness decrease (*FrDec*), as well as the reference attitude (*Ref*). The five conditions were tested in a between-subjects design.

Participants are assigned to one condition. If the condition is *Ref* then participants were asked to view seven reference videos and rate each of them by answering 20 questions such as *"in your opinion, the behavior of the virtual character is assertive?*. For the other conditions (*DomInc, DomDec, FrInc, FrDec*), participants view and compare seven pairs of videos: reference video vs. comparison video, then they rate the behavior of ECA in the comparison video. An example of comparison question is: *"compared to the reference video (left), the behavior of the virtual character in the comparison video (right) is assertive?*. In addition to the 16 dependent variables used in the first experiment, we added four new variables: *dominant*, *submissive*, *friendly*, and *hostile*. All answers were on a 5-point labeled Likert scale (1 ="strongly disagree", 2 = "partially disagree", 3 = "neutral", 4 = "partially agree", and 5 = "strongly agree"). The synthesized speech was identical for each pair of videos.

We are aware the voice shows also attitude change [Janssoone, 2015]. But, since, on the one hand, we have not focused our work on acoustic feature of attitude changes, and on the other hand, most voice synthesizers do not model attitude changes, we decide to use neutral voice for each video of each condition. We conduct comparison studies where all the videos use the same voice synthesizer. After viewing the current pair of videos and answering the questions related to the agent's behavior, the next pair of videos is automatically displayed. We show questions and videos according to a latin square design. We build on the same hypotheses as in the first experiment (cf. Section 7.6.3).

6.3.2 Results

A total of 90 participants have been recruited from the Crowdflower platform (18 for each condition); only 20% were female, 50% had a master level and all participants were Europeans or Americans (34% Spanish, 20% German, and 18% French). We present the results in the same way as for the first experiment: we start by plotting the circular profile of the agent as well as its vector length. Then, we investigate the differences in the ECA perception between the comparison video and the reference video. Finally, we compute performance scores (recall, precision and F-measure) for attitude recognition.

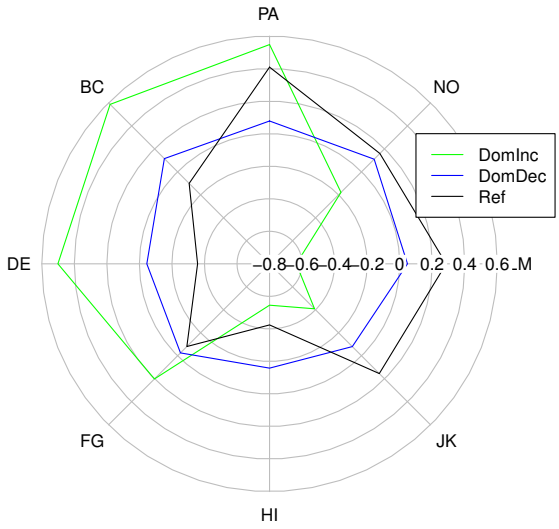


Figure 6.10 – Plotting octant scores on the IPC for dominance increase and decrease.

Circular Profiles

Figure 6.10 and 6.11 show the octant scores respectively for dominance and friendliness variations. From these graphs, we can draw the following conclusions:

- For *DomInc*, the ECA is perceived as **more dominant**, **more hostile**, **less friendly**, and **less submissive** compared to the agent in *Ref*;
- For *DomDec*, the profile of the ECA is circular: participant gives the same value for all octants;
- For *FrDec*, the agent is evaluated as **more hostile** and **more dominant** compared to the agent in *Ref*;
- For *FrInc*, the ECA is perceived as **less friendly**, and **less dominant** compared to the agent in *Ref*.
- ECA in the reference video is perceived as **friendly**, **dominant**, **not hostile** and **not submissive**.

Vector Scoring

Table 6.1 gives the angle and vector length for all conditions. For *FrInc* and *Ref*, the vector angles of both agents are situated in the NO octant (friendliness-dominance region). For *DomInc*, the vector angle of the agent is in the BC octant (dominance-hostility region) with an intense attitude amplitude (vector length = 0.078) compared to the other conditions. For both *DomDec* and *FrDec*, the vector angles of agents are in the dominance region (PA). However, the agent communicating a dominance decrease has a small vector length compared to the other conditions.

Result significance

To assess whether a significant difference exists between the perception of the reference

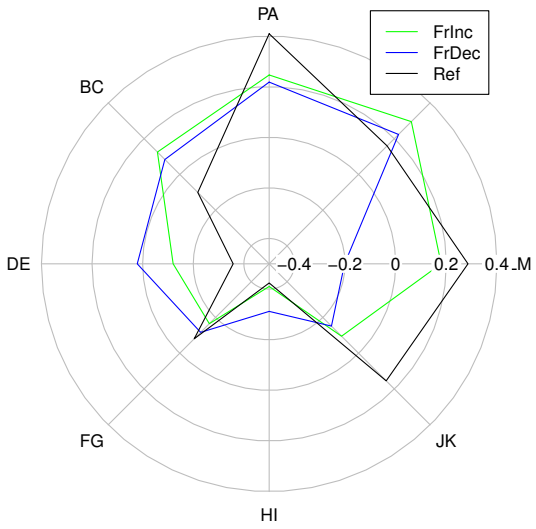


Figure 6.11 – Octant scores for friendliness increase and decrease.

	DomInc	*DomDec*	*Ref*	*FrInc*	*FrDec*
Vector angle	144.08	83.81	67.56	58.62	80.01
Vector length	0.12	0.64	0.86	0.87	0.88

Table 6.1 – Angle and vector length of each condition.

and of the comparison videos, we conducted a Wilcoxon test. In appendix B, we report, for each condition, the mean, standard deviation of independent variables, and distribution of participants' answers over these variable. The revealed differences are illustrated in Figure 6.12. For *FrInc* and *DomDec* no significant differences have been detected. For *DomInc*, the agent is perceived as:

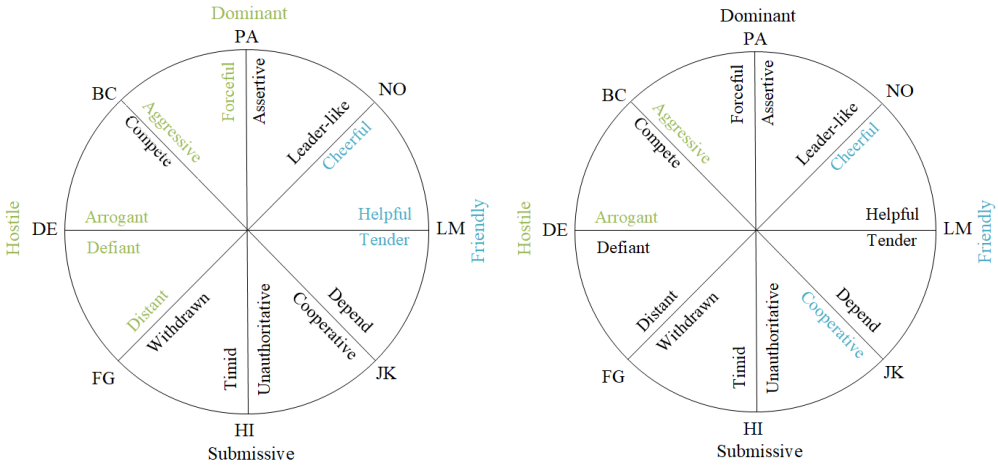
- **More dominant**: *aggressive* ($p < .001$), *forceful* ($p = .001$) and *dominant* ($p = .002$) compared to the agent in the reference video, therefore **H.Dom** is supported;
- **More hostile**: *arrogant* ($p < .001$), *defiant* ($p < .001$), *distant* ($p = .005$) *hostile* ($p < .001$) compared to the agent in the reference video;
- **Less friendly**: *helpful* ($p < .001$), *cheerful* ($p = .004$), *tender* ($p = .01$), and *friendly* ($p < .001$) compared to the agent in the reference video.

Finally, for *FrDec*, the ECA is evaluated as:

- **More dominant**: *aggressive* ($p = .003$);
- **More hostile**: *arrogant* ($p = .01$), *hostile* ($p = .01$), thus **H.Hos** is accepted;
- **Less friendly**: *cooperative* ($p = .003$), *helpful* ($p < .001$), and *friendly* ($p = .001$).

Information Retrieval Metrics

The problem of evaluating the attitude of an ECA expressing a given attitude variation can



Dominance increase (a) **Friendliness decrease (b)**

Figure 6.12 – Differences in the ECA perception between the comparison video and the reference video. The blue color indicates that the variable is evaluated as being more expressed in the reference video; the green indicates that the variable is evaluated as being more expressed in the comparison video; and the black denotes no difference between the reference video and the comparison video.

be viewed as multi label classification, where the predicted class (label) can be one or more among the four attitude variations (*DomInc*, *DomDec*, *FrInc*, and *FrDec*). As such, the four tested conditions can be evaluated based on the classical measures: recall, precision and F-measure. Table 6.2 gives recall, precision, and F-score for the four attitude variations. The best results are achieved for *DomInc*. Table 6.3 reports the distribution of the predictions over the actual conditions. A cell where actual$=A$ and predicted $=B$ gives the number of videos actually expressing A and evaluated as expressing B. Based on Table 6.3, we observe that the ECA in *Ref* is evaluated as **friendly**, **dominant**, **not hostile** and **not submissive**.

	Dominance		Friendliness	
	Increase	Decrease	Increase	Decrease
Recall	63%	18%	39%	26%
Precision	37%	44%	31%	20%
F-measure	47%	26%	35%	22%

Table 6.2 – Recall, precision, and F-measure for the multimodal model.

6.4 Discussion

The experiment conducted to evaluate the *Sequential Attitude Planner* has led to some interesting findings. Table 6.4 recapitulates the most significant conclusions. We specifically focus on the correlation between attitude variation and attitude perception. For each combination, we report whether it is statistically significant (column Sign.), if it coincides

		Predicted			
		DomInc	*DomDec*	*FrInc*	*FrDec*
Actual	*DomInc*	**62%**	31%	38%	**41%**
	DomDec	13%	18%	18%	6%
	FrInc	**35%**	14%	**39%**	15%
	FrDec	**59%**	25%	16%	**25%**
	Ref	**43%**	12%	**39%**	14%

Table 6.3 – Distribution of the predictions over the actual conditions.

with the circular profile of the agent (column IPC) and we also indicate the direction of the relationship (column Direct.).

	Dominance			Submission			Friendliness			Hostility		
	IPC	Stat.	Direct.	IPC	Stat.	Direct.	IPC	Stat.	Direct.	IPC	Stat.	Direct.
DomInc	✓	✓	↗	✓	x	↘	✓	✓	↘	✓	✓	↗
DomDec	✓	x	↘	✓	x	↗	✓	x	↘	✓	x	↗
FrInc	✓	x	↗	✓	x	↘	✓	x	↗	✓	x	↘
FrDec	✓	✓	↗	✓	x	↘	✓	✓	↘	✓	✓	↗

Table 6.4 – Relationships between attitude variations and attitude perception.

The two hypotheses **DomInc** and **FrDec** are supported. As in the first evaluation study, the agent in the reference condition is perceived as friendly. The hypothesis **H.Ref** is rejected. The agent expressing friendliness increase is evaluated as equivalent to the agent in the reference video, that is friendly, therefore, **H.FrInc** is rejected. But we can notice that, since by default (*FrRef*) the virtual agent appears friendly, it appears also friendly in *FrInc*.

Unlike our first study, our **H.sub** hypothesis is not validated. Chollet and colleagues [Chollet et al., 2014b] found similar result with their virtual recruiter displaying dominance decrease. This result can be related to the context of the interaction where the agent plays the role of a job recruiter. In such a context, the recruiter tends to controls the interaction and therefore appears naturally dominant and not submissive. Furthermore, the agent in *DomDec* is perceived as more hostile and less friendly compared to the agent in the reference video (*DomRef*) while in the first study it is perceived as more friendly and less hostile. This change in perception confirms the importance of the interaction context that can alter the perception of an attitude.

Another results that may be related to the role of the agent is the halo effect between the dimensions of friendliness and of dominance: increasing friendliness leads to the perception of dominance increase. However, in the first study friendliness increase was correlated with dominance decrease. There is a compensation effect.

An attitude dimension is represented by two symmetrical poles (dominance/submission, friendliness/hostility). We were expecting a negative relationship between the two poles of an attitude dimension (an increase of a given pole would result in a decrease in the perception of the opposite pole). This assumption is statistically significant for friendliness/hostility: when the agent is evaluated as more hostile it is also perceived as less

friendly, and vice-versa. For the other dimension, there is a strong trend: when the agent is evaluated as more dominant it is also perceived as less submissive, and vice-versa.

6.5 Conclusion

In this chapter, we presented the “Sequential Attitude Planner" that generates the behavior of the agent according to its communicative intentions as well as the attitude variation it should express. First, the planner converts the communicative intentions into a sequence of non-verbal signals. Then, from the database of sequences expressing the input attitude variation, the planner chooses the most representative sequence of the input attitude variation that is also the most similar to the communicative sequence (in term of non-verbal behaviors). Both sequences are then combined to compute the final sequence conveying both intentions and attitude variation.

We conduct an evaluation study where we compare videos of the agent communicating with and without an attitude. We can conclude from this study that our attitude planner allows the ECA to express a variation of attitudes, in particular *dominance increase* and *friendliness decrease*. The non recognition of the variation *dominance decrease* could be caused by the interaction context, here the role of the agent. The friendliness perception of the agent in the reference conditions seems to affect the recognition of the variation *friendliness increase*.

Take home

- Attitude planner allows the ECA to express both attitude variation and communicative intentions.
- Attitude planner allows the ECA to express: *dominance increase* and *friendliness decrease*.
- High correlation between *dominance increase* and *friendliness decrease*.
- The role of the agent could influence the perception of its attitude.

Chapter 7

Generative Model of Agent's Behaviors in Human-Agent Interaction

Contents

7.1 Related works **83**
7.2 Corpus **84**
7.3 Neural Networks and LSTM **87**
7.3.1 Neural Networks: Overview 87
7.3.2 Recurrent Neural Networks: LSTM 88
7.4 LSTM Model **89**
7.4.1 Prediction Model 89
7.4.2 Evaluation 91
7.5 Architecture **92**
7.5.1 EyesWeb: User's Behavior Analysis 93
7.5.2 Flipper: Dialog Management 93
7.5.3 GRETA-VIB 95
7.5.4 BehaviorPrediction: Agent's Behavior Prediction 96
7.6 Evaluation **97**
7.6.1 Independent Variables 97
7.6.2 Measures 97
7.6.3 Hypotheses 98
7.6.4 Protocol 100
7.7 Results **100**
7.8 Conclusion **102**

IN human interaction, humans adapt and adjust their behavior according to the behavior of their interlocutors [Burgoon et al., 2010]. For example, a listener nods for indicating his agreement with the speaker, he gazes at the same object or smiles in response to the interlocutor's smile. In light of this, we aim to model an Embodied Conversational Agent (ECA) able to adapt its behavior according to the user's behavior. Nonverbal behaviors play an important role for maintaining engagement between humans and agent [Fong et al., 2002, Arai and Hasegawa, 2004, Breazeal., 2004, Woolf and Burleson, 2009]. This is why we are particularly interested in adapting dynamically the agent's nonverbal behaviors to those of its interlocutor.

To adapt the agent's behavior according to the user's one, we take advantage of the recent advances in the domain of neural networks, specifically a popular type of networks called LSTM. This approach simultaneously encompasses the sequentiality and temporality of non-verbal behavior over time. The designed model adopts a user-in-the-loop approach to constantly predict the behavior of the agent in response to the user's behavior. It takes as input both, the user's and the agent's past behavior and predicts the next agent's behavior. More precisely it predicts agent's *smile*, *head movements*, and *gaze*.

To integrate and evaluate our LSTM model called IL-LSTM (Interaction Loop LSTM), we create an interaction system where the agent interacts in real-time with a human user. The system takes as input data from the user, computes what the agent has to say as well as the corresponding animation. It is built upon four modules: (1) User's behavior detection and analysis based on multimodal analysis software EyesWeb [Volpe et al., 2016]. (2) Dialogue management: we have used the dialog manager Flipper to define the dialogue rules (turn taking and verbal content) for the virtual agent [van Waterschoot et al., 2018]. (3) Agent's behavior prediction based on our IL-LSTM (Interaction Loop LSTM) model to predict the behavior of the agent for the next frame taking as input the behavior of both agent and user, during the past frames. (4) Behaviour generation from the predicted agent's behavior using the GRETA-VIB platform [Pecune et al., 2014]. To our knowledge, our model is the first attempt to produce, in real time, smile, head movements and gaze direction for virtual agent driven from both agent and user's smile, head movements, and gaze direction as well as agent communicative intentions. The results showed that users were indeed more satisfied by their interaction with the agent when it adapted its behavior.

In this Chapter, we present an IL-LSTM model that continuously generates the agent's behavior that is both responsive to the user's behaviors and is consistent with the agent's intentions. First, we present an overview of related works 7.1. Then, in Section 7.2 we describe the database we used to train our model. In Section 7.3, we present an overview of the neural networks. Our IL-LSTM model is described in Section 7.4. Then, we present, in Section 7.5, the architecture of the interactive system implementing the IL-LSTM model. The experimental protocol used to evaluate our interactive system as well as the obtained results are presented in Section 7.6.

7.1 Related works

In this work, we focus on adapting the agent's nonverbal behavior according to user's behavior. In this Section, we present the related works to our problematic.

In [Woolf and Burleson, 2009], authors used a rule-based approach to adapt the facial expression of a virtual tutor according to the student's affective state (e.g., frustrated, bored, or confused). For example, if the student is sad, respectively delighted, the tutor might look sad, respectively pleased. Results showed that when the virtual tutor adapts its facial expression in response to the student's one, the latter maintained higher levels of interest and reduced boredom when interacting with the tutor. However, in this work the agent does not react Instantaneously to the student's affective state but it responds only when the student has finished his speaking turn.

In [Gordon et al., 2016], authors rely on reinforcement learning algorithms to learn how to properly respond to child's affect (valence). Results showed a significant increase in child's valence when the robot adapted its facial expression to the child's one, compared to a non-adaptive robot.

Other works focus on modeling the agent's facial expression in dyadic interaction. In [Huang and Khan, 2017], authors used Generative Adversarial Networks (GANs) [Goodfellow et al., 2014] to model the facial expressions of interviewer within interviewer-interviewee interactions. This model is decomposed into two steps: (1) generating face sketches of the interviewer taking as input facial expressions of the interviewee; (2) synthesizing complete face images conditioned on the face sketches generated in the first step. Results showed that the facial expressions in the generated interviewer face images reproduce appropriate emotional reactions to the interviewee behavior. A very similar work has been presented in [Nojavanasghari et al., 2018].

Feng and his colleagues relied on Neural Networks to continuously predict facial expressions during human-human interaction [Feng et al., 2017]. They used LSTM to predict the facial expression from Skype conversations involving pairs of persons engaged in conversations. The goal is to predict the next facial keypoints (i.e., landmarks) of one person taking as input the previous facial keypoints of both partners, unlike other works that infer the facial expression of one partner only from the facial expression of the other partner.

Table 8.1 gives a comparison of the works presented above as well as our model. Each entry is characterized with a set of criteria: the algorithm used to predict the nonverbal behavior, the considered features, instantaneous (Inst.) prediction and generation, the consideration of interaction loop (whether the model takes into account both partners' behavior in a dyadic interaction), and generation (whether the model has been integrated into a virtual agent to predict the agent's behavior in real time). Also, we indicate the contribution of each model to the human-agent interaction. Traditionally the rule-based approach has been popular [Arai and Hasegawa, 2004, Breazeal., 2004, Woolf and Burleson, 2009]. Very recently the trend turns to using neural networks, such as LSTM and GANs that have been particularly efficient for modeling and generating nonverbal behaviors in dyadic interaction. These kind of models naturally encompass the dynamics of the behaviors. This allows for predicting continuously (i.e. on a frame by frame basis) the

Ref	Algorithm	Feature types	Inst.	Int. loop	gen.	Results
[Woolf and Burleson, 2009]	heuristic strategies	facial expression	x	x	✓	decrease student's boredom
[Gordon et al., 2016]	reinforcement learning	facial expression	x	x	✓	increase in child's valence
[Huang and Khan, 2017]	GAN	facial expression	✓	x	x	
[Feng et al., 2017]	LSTM	facial expression, head pose	✓	✓	x	
[Nojavanasghari et al., 2018]	GAN	facial expression, head pose	✓	x	x	
Our model	LSTM	Smile, head movements, gaze direction	✓	✓	✓	increase user satisfaction

Table 7.1 – An overview of works related to adaptation model.

behaviors, unlike the rule-based approach that generate the agent's behaviors for a certain time windows. As we can observe from Table 8.1, in most works the agent's facial expression exclusively depends on the user's facial expression and ignores the interaction loop between the agent and the user. Feng's work is the first attempt to take into account this interaction loop, i.e., predicting the behavior of one interaction partner taking as input the behavior of both interaction partners.

Despite being a significant contribution, the works of Feng and those based on neural models alike, have not been explored for real-time generation of the agent's facial expression. They settled for learning the models from human interactions. To continue these efforts, we design a model that goes beyond learning the facial expressions to their generation into an ECA platform. Thus, we would be able to predict, in real time, the agent's smile, head movements, and gaze direction that are responsive to the user's behavior, allowing their behaviors to appear as a dynamic process in an interactive loop. The other novelty of our model is to include the agent's communicative intention (e.g., emotion) along with the other behaviors related to the agent's adaptation.

7.2 Corpus

In the context of the ARIA-VALUSPA project, a corpus called NoXi (NOvice eXpert Interaction database) [Cafaro et al., 2017] has been collected. NoXi focuses on knowledge sharing between an expert and a novice discussing about a given topic (e.g., sports, politics, videogames, travels, music, etc.). The recording was conducted in three different countries: France, Germany and UK. In the context of this thesis, we participated to the creation of this corpus by collecting and annotating the French part of NoXi. Figure 7.1 reports the most recurrent topics discussed in each of the three countries.

NoXi is a corpus of spontaneous and screen-mediated face-to-face interactions. Expert and novice were in two different rooms and interacted through a large screen as illustrated

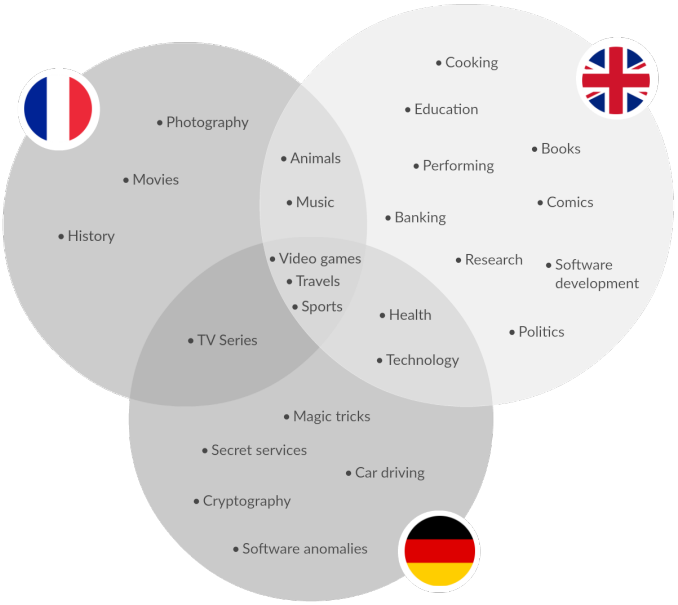


Figure 7.1 – The most frequently-discussed topics in NoXi corpus.

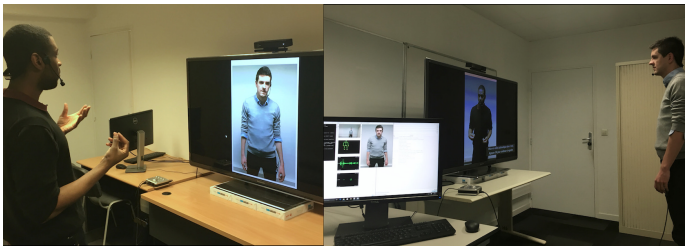


Figure 7.2 – Expert-novice interaction.

in Figure 7.2. NoXi is publicly available through a web interface[1]. The dataset contains over 25 hours of dyadic interactions spoken in multiple languages (mainly English, French, and German).

For training our model, we use the French part of NoXi database which is composed of 20 sessions. The total duration of all these sessions is 6 hours and 52 minutes (618000 frames). We automatically extracted the facial expressions and semi-automatically detected the conversation states of interactions:

- **Conversation state (semi-automatic)**:
  The conversation state corresponds to which interlocutor is speaking. We have defined four states: both interlocutors speak (both), expert speaks (expert), novice speaks (novice), or no one speaks (none). Based on both expert and novice's voice activity detection, the conversation states of the interaction has been automatically annotated (see Figure 7.3). As the voice activity presents some noise, we manually adjust the automatic annotation.

[1] https://noxi.aria-agent.eu/

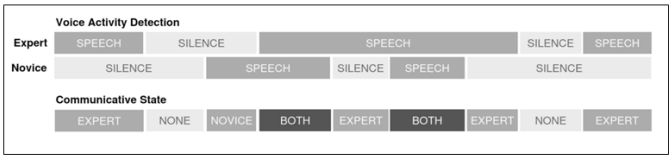


Figure 7.3 – Example of annotations of the conversation states detected from Expert and Novice's voice activity.

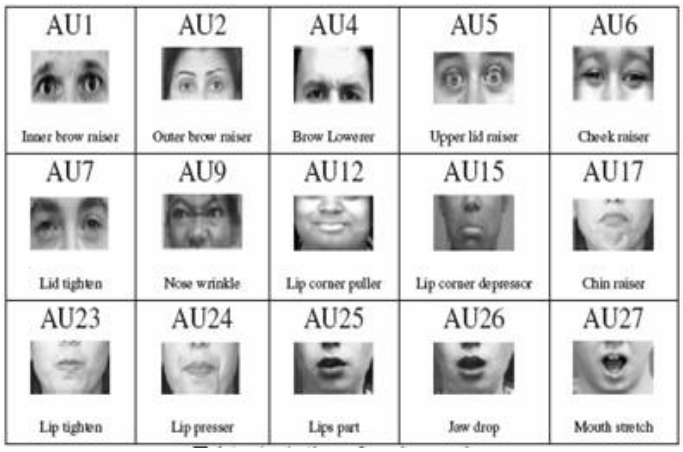


Figure 7.4 – Action Units corresponding to the activation of different facial muscles.

Table 7.2 gives the percentage of each conversation state. We observe that the expert speaks more (57%) than the novice (24%). This was expected as the expert controls the discussion topic and speaks more to explain his topic.

None	Novice	Expert	Both
14.05%	24.48%	57.51%	3.96%

Table 7.2 – Percentage of each conversation state in NoXi database.

- **Facial expression extraction (automatic)**
  Several works focus on facial expression extraction [Mackenzie et al., 1985, Sariyanidi et al., 2015, Wang et al., 2018, Murphy-chutorian et al., 2009]. We use the open-source tool OpenFace [Baltrusaitis et al., 2016] to extract facial expressions of both expert and novice[2]. OpenFace allows the extraction of head pose and rotation, gaze, and facial Action Units. Action Units (AUs) represent the movements of facial muscles classified according to the FACS (Facial Action Coding System) taxonomy [Ekman and Friesen, 1976]. Figure 7.4 shows examples of AUs[3]. For example, AU12 is estimated based on the activation of the zygomaticus major muscle.

  We are interested by predicting agent's facial expression, in particular, gaze, smile and head movement. We use OpenFace to extract the information related to these behaviors; i.e., AU12 (smile), head rotation as well as the gaze:

[2] https://github.com/TadasBaltrusaitis/OpenFace
[3] https://www.ecse.rpi.edu/~cvrl/tongy/aurecognition.html

- The AU12 intensity (varies from 0 to 5)
- The AU12 activation (0 absent, 1 present)
- Head rotation: rotation is in radians around X,Y,Z axes
- Gaze direction: gaze direction of both eyes in radians.

More details about these features are given in https://github.com/TadasBaltrusaitis/OpenFace/wiki/Output-Format

In the next section, we present an overview of the neural networks.

7.3 Neural Networks and LSTM

Since the beginning of this decade, a hype has occurred around the domains of Deep Learning and Neural Networks. Mainly motivated by the dramatic increase in computing power capacities (parallel computing, GPUs), Neural Networks have gained back attention among academics and technology giants, like Google and Facebook.

Neural Networks have re-established themselves as a must in many domains, especially when it comes to machine learning tasks involving either complex or unstructured data, like image recognition, text/speech processing and translation, etc. A wide variety of networks exist, differing from each other by their structure and the learning task they have been designed for. Network structure can be as simple as a perceptron or a Feed-forward Network, generally used for simple classification tasks. Other structures have been designed, and are used as basic units to build customized networks for specific tasks, such as Convolutional Networks for image processing, Recurrent Networks for time-aware tasks, Autoencoders specifically relevant for text and speech processing, etc.

The next Section is intended to shed light on Recurrent Neural Networks (RNNs) with a focus on a major RNN variation called LSTM. For the sake of brevity, and to avoid too much dilution, this overview is intentionally limited to the essentials, despite its obvious relevance to our topic.

7.3.1 Neural Networks: Overview

To better explain and understand RNNs, one should first understand the logic behind Neural Networks, and especially, the simplest type of them: Feed-forward Neural Networks (FNNs). Inspired by the biological brain, an FNN is composed of a set of processing units (nodes), in addition to communication (input/output) nodes. FNNs are named after the way these nodes are interconnected: information are fed straight through, neither loops nor backtracking are allowed in this type of architecture (Figure 7.5).

Input examples are fed to the network then transformed gradually until reaching the final (output) stage. At each intermediary layer, an operation is performed on the input values and fed to the next layer. Usually, this operation is a linear combination of inputs to which is added some post processing (e.g. activation functions). FNNs are generally trained on large amounts of data. They have demonstrated very satisfying results for

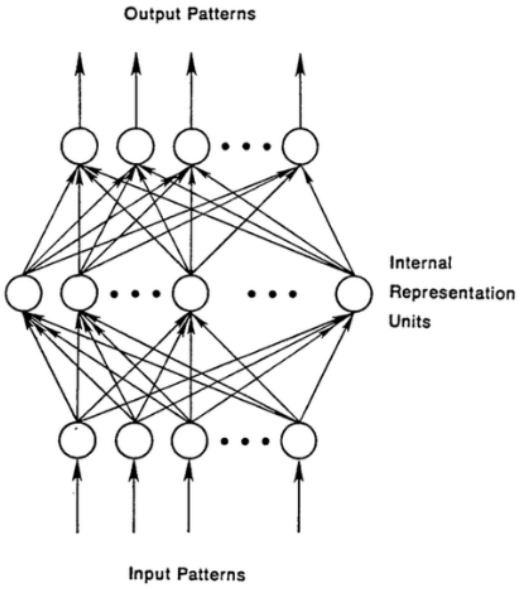


Figure 7.5 – A simple neural neural network architecture.

many supervised and unsupervised classification problems. In supervised learning, the model parameters are adjusted iteratively to minimize an objective function, called Loss. Loss can be measured by a distance value or an error rate empirically calculated on the learning examples.

7.3.2 Recurrent Neural Networks: LSTM

RNNs are family of Neural Networks, specifically designed to deal with sequential and/or temporal data. In the diversity of real-life data, sequentiality is a frequent pattern; e.g., written text, speech, and even images can all be considered as sequences of small amounts of data (words, pixels, etc.). Unlike FNNs, an RNN takes as input not only the current input example but also what it has perceived in the past time (a sort of "memory"). Since RNNs rely on the concept of memory, the analogy, one more time, can be made with the biological brain. Concerning the network structure, RNNS are distinguished from FNNs by their feedback loop connected to their past decisions. This is why it is often said the RNNs have memory. This kind of networks is mainly used when "context" is important, i.e., decisions from the past can influence the current ones. A common example is text where words should be analyzed within their context and with knowledge of the preceding ones.

Hochreiter and Schmidhuber introduced a particular recurrent network called Long Short-Term Memory (LSTM) Hochreiter1997 [Hochreiter and Urgen Schmidhuber, 1997]. While an RNN makes use of the near-past events (context), an LSTM has been designed to deal with frequent events occurring in either near or far-away past. In theory, RNNs are absolutely able to handle long-term memory but in practice, they do not seem to succeed in learning the parameters to solve the problem [Bengio et al., 1994].

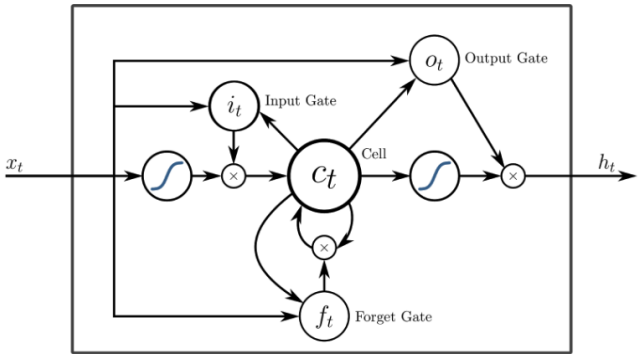


Figure 7.6 – LSTM cells.

The basic element of LSTM is the so-called memory cells, along with a couple of elements, called gates, that are recurrent and that control how information is being remembered and forgotten (Figure 7.6). In the world of machine/deep learning, LSTMs have a massive commercial success, especially in the domains of speech recognition and machine translation[Tian et al., 2017].

LSTM are achieving remarkable results across various domains, where sequentiality and temporality are important factors. Encouraged by these achievements, we choose to apply LSTM for behavior modeling, and specifically, for predicting the next ECA's behavior based on the 20 past facial expressions of ECA's and of user's. The next Sections give more details on the methodology we followed and the achieved results.

7.4 LSTM Model

In this section, we present our LSTM model that we call IL-LSTM for Interaction Loop LSTM.

7.4.1 Prediction Model

We aim to continuously update the agent's expression that is both responsive to the user and is consistent with the agent's communicative intentions. Thus, the agent's behavior is computed to convey its communicative intentions and to adapt to user's nonverbal behavior. That is, the computational model that controls the behaviors of the agent takes as input the multimodal behaviors that convey the different communicative intentions of the agent as well as the user's nonverbal behaviors displayed simultaneously. For this, our model was designed to use the multimodal behavior of both, the agent and the user (head rotation, gaze, smile) as well as the conversation state (Who is speaking?) of the interaction as input to predict head rotation, gaze and smile of the agent.

Our model will be tested with a virtual agent playing the role of an expert of video games that describes a video game exhibit to a user that is going to visit it. We train and evaluate our model using the NoXi database where we choose the novice to play the role

of the user (i.e. he has to learn about the exhibit), and we choose the expert to play the role of the agent (i.e. it has knowledge about the exhibit).

Our data includes two types of features: numerical (smile intensity, head rotations, and gaze directions) and categorical (conversation state with labels like "none", "expert", etc.). Before presenting our model, we prepare the input data for neural networks. Neural networks work internally with numerical data which requires the conversion of categorical data into numerical form. One-hot encoding or categorical encoding is a widely used transformation technique for categorical data. One-hot encoding represents each category as an all-zero vector with a 1 in the place of the category index [Bengio, 2012, Chollet, 2017]. We use One-hot encoding to encode conversation state. Thus, the four conversation states are encoded by the list: $[1, 0, 0, 0, 0]$ (none), $[0, 1, 0, 0, 0]$ (expert), $[0, 0, 1, 0, 0]$ (novice), $[0, 0, 0, 1, 0]$ (both).

Our data is heterogeneous: different features take different ranges. For example, AU intensities vary between 0 and 5, whereas head rotation ranges from 0 to 2π. Such data must be normalized before feeding it into a neural network relying on feature-wise normalization [Bengio, 2012, Chollet, 2017]. Thus, all numerical features vary within the range of 0 and 1.

The LSLM model allows us to model both sequentiality and temporality of non-verbal behavior. It takes as input a sequence of features observed during a sliding window of n seconds (equivalent to m frames) and give the prediction for the next frame. Our model IL-LSTM has a single LSTM hidden layer to extract features from the input sequence, followed by a fully connected layer to interpret the LSTM output, followed by four output layers used for predicting the four output variables (see Figure 7.7).

The IL-LSTM includes a regression task for smile intensity, gaze, and head rotation prediction and binary classification task for smile activation prediction. We used cross entropy as loss function for the classification task and mean squared error (MSE) loss for the regression task. As activation function, we used *sigmoid* for making binary prediction and *relu* (rectified linear activation unit) for regression prediction. When it comes to the training, we used mean absolute error (MAE) for evaluating the regression scores and accuracy for the classification predictions. MAE represents the absolute value of the difference between the predicted scores and the targets, whereas accuracy indicates the fraction of the correct smile predictions. For evaluation purpose, 80% of data were used for training the model, 10% for validation, and 10% for testing.

The prediction model takes as input a sequence composed of the last n frames (one frame is equal to 0.04 second) of agent's and user's features and predicts the agent's behavior for the next frame. To help fixing the sequence length of the input, we vary n from 0 to 50; that is we vary the number of past frames we consider to predict the agent's behavior at the next frame. We choose 20 frames (for a total duration of 0.8 seconds) since the model loss stops decreasing after this value.

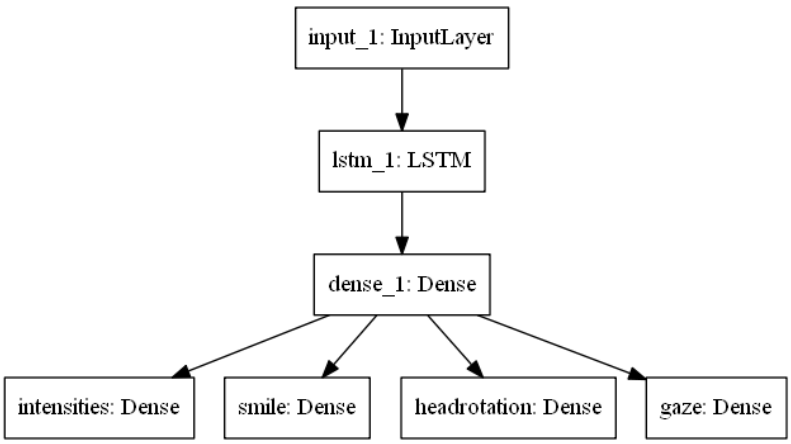


Figure 7.7 – Our neural network graph with multiple outputs.

7.4.2 Evaluation

We measured the performance of our regression models using the root-mean-square error ($RMSE$) and the coefficient of determination (R^2). These are widely used measures for regression problems. The $RMSE$ is defined in Equation 7.1, where y_{gt} are the ground-truth observed variables, y_{pred} are the predicted values, and N is the number of data samples:

$$RMSE = \sqrt{\frac{1}{N}\Sigma_{i=1}^{n}(y_{gt} - y_{pred})^2} \tag{7.1}$$

The coefficient of determination R^2 is defined in Equation 7.2, where y_{gt}, respectively $\overline{y}$, are the observed target variables, respectively their mean values, and y_{pred} are the predicted values. It is based on the ratio between the mean squared errors (MSE) and the variance in y values (denominator). This measure reflects the improvement over the average-baseline predictor ($y_{pred} = \overline{y}$). When an average-baseline is used, the value of R^2 is close to zero. Thus, an under-performing model may yield negative values.

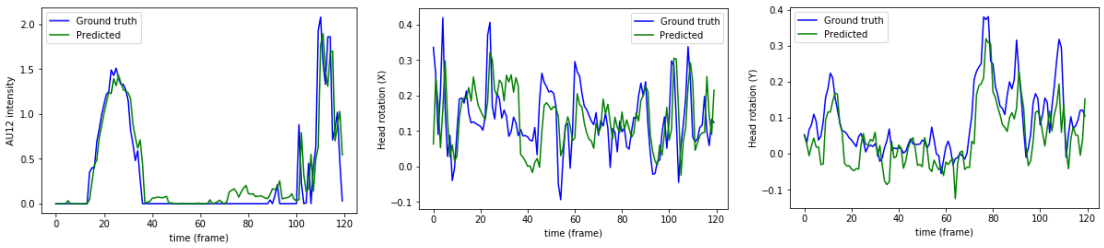
$$R^2 = 1 - \frac{\Sigma_{i=1}^{n}(y_{gt} - y_{pred})^2}{\Sigma_{i=1}^{n}(y_{gt} - \overline{y})^2} \tag{7.2}$$

For comparison purpose, we use a baseline predictor that returns the average of each agent's behaviors calculated on the past 20 frames. Table 7.3 reports the performance of our model in comparison to the baseline. We observe that our model significantly outperforms the baseline model. When it comes to the performance of predicting each modality, we observe that our model achieves better performance for the prediction of smile intensity (R^2=0.809) than head rotation (R^2=0.481), and gaze direction (R^2=0.552).

Figure 7.8 shows ground truth and predicted agent's smile and head rotations. We observe that predicted smiles appear approximately at same time as the ground truth and with the same intensity. Regarding head rotations, our model is able to predict head rotation similar to the ground truth in term of both head rotation direction and value.

		Smile intensity	Head rotation	gaze
Baseline	R^2	-62.985	-1.047	-0.015
	$RMSE$	0.886	0.566	0.784
Our model	R^2	0.809	0.481	0.552
	$RMSE$	0.275	0.082	0.068

Table 7.3 – Performance of our model in comparison to the baseline.



Smile **Head rotation (x)** **Head rotation (y)**

Figure 7.8 – Plots of ground truth and predicted agent's smile and head rotations (x and y) through time.

In this section, we present our IL-LSTM model that takes as input the behavior of the agent and of the user to predict the agent's facial expression for the next frame. In the next Section, we present an interactive system in which our IL-LSTM model is used to predict in real time the behavior of the agent.

7.5 Architecture

We have created a system where the agent interacts in real-time with a human user. It takes as input the user's behavior and computes the agent's behavior (speech and non-verbal behavior). It is built upon four modules as illustrated in Figure 7.9:

- User's behavior analysis: we rely on EyesWeb to detect the user's behavior.
- dialog management: we use Flipper to define the dialog rules (turn taking and verbal content) for the virtual agent.
- Agent's behavior prediction (IL-LSTM): this module uses our LSTM model to predict the behavior of the agent for the next frame taking as input the behavior of both, the agent and the user, during the last 20 frames.
- Behaviour generation (GRETA-VIB): the predicted agent's behavior is fed to the agent using the GRETA-VIB framework.

In the following, we describe each of these modules.

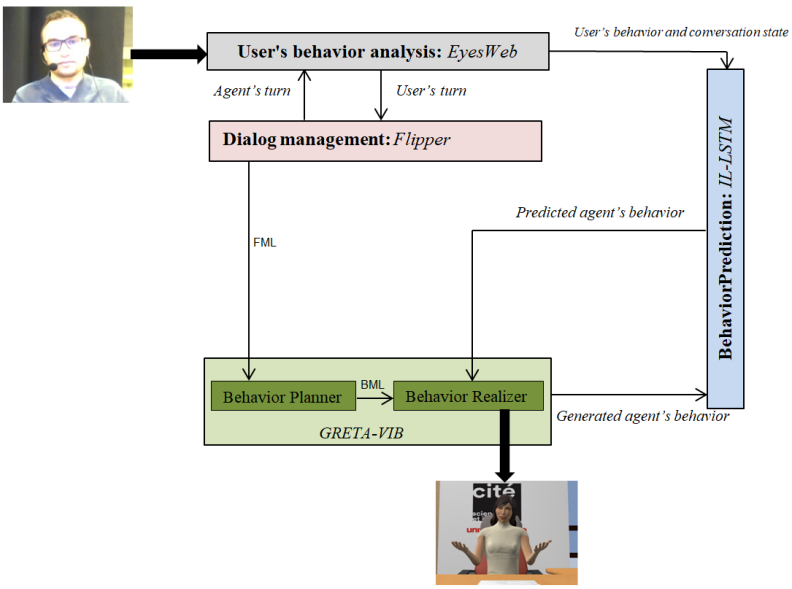


Figure 7.9 – Architecture of our system.

7.5.1 EyesWeb: User's Behavior Analysis

EyesWeb XMI is an open source platform for recording and analyzing human behavior in real-time [Volpe et al., 2016]. EyesWeb can be connected to different input devices like camera, Kinect, microphones, etc. It supports real-time synchronized recording of multimodal channels. It also includes several libraries like OpenFace. Using EyesWeb, we extract the user's facial expressions, head movements, and voice activity. Figure 7.10 shows the EyesWeb interface.

7.5.2 Flipper: Dialog Management

During human-agent interaction, the agent must choose the next dialog act depending on the evolution of the interaction with the user. For example, the agent starts the interaction by greeting the user, after the user responds, the agent will introduce itself and asks the user about his name and so on. To select the most appropriate dialog acts (including how it is instantiated in terms of speech and communicative intentions) of the agent, we use Flipper [van Waterschoot et al., 2018]. Flipper is an open-source library developed to specify dialog rules for virtual agents.

In our system, Flipper manages the agent's speaking turn and sends messages (*Agent's turn*) to EyesWeb to indicate when the agent starts and finishes speaking. For that, EyesWeb sends to Flipper the voice activity of the user (when the user starts and finishes speaking). Based on the dialog step, Flipper selects the next agent's speech (augmented

Figure 7.10 – EyesWeb interface. Relying on OpenFace, EyesWeb extracts the user's facial expressions.

```
<is name="informationstates">{
        "user": {
            "voiceActivity": 1,
        },
        "agent": {
            "fml": {
                "template": "",
                "parameters": {}
            },
            "isTalking": false
        },
        "dialogue": {
            "step": {
                "current": "greeting",
                "next": ""
            }
        }
    }
</is>
```

Figure 7.11 – Information state of our system.

with communicative intentions and defined in a file following the Functional Markup Language FML format [Heylen et al., 2008]).

Flipper defines two concepts, namely the information state and templates to manage the dialog of a virtual agent. The information state stores the agent's knowledge of the interaction (e.g., conversation state). The templates describe the preconditions and effects of the dialog rules. Preconditions indicate the conditions that should be checked for triggering a given dialog act (e.g., greeting, asking user's age, describing a video game). Effects are the associated updates to the information state. In our system the information state contains information about the agent, user, and dialog state as shown in Figure 7.11.

We rely on the "current" dialog act of the agent to launch the "next" one. For example, after asking the user's age, the next dialog act asks for user's country. The template corresponding to this example is shown in Figure 7.12. This template triggers two events:

- sending the FML file (dialog act: ask user's age) to the "behavior planner";
- changing the information state: the "next" step in dialog will be "ask country".

```
<template id="ask_age" name="ask age">
    <preconditions>
        <condition>
            is.dialogue.step.current == "ask_age"
        </condition>
    </preconditions>
    <effects>
        <assign is="is.states.agent.fml">"ask_age"</assign>
        <assign is="is.dialogue.step.next">"ask_country"</assign>
    </effects>
</template>
```

Figure 7.12 – Example of Flipper template.

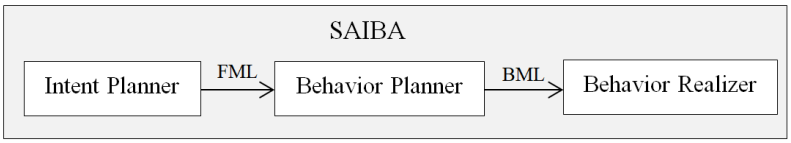


Figure 7.13 – SAIBA architecture.

The state of a next dialog (in the example “asking country”) is changed from “next” to “current” when the user finishes answering the current dialog (in the example “asking the age”).

7.5.3 GRETA-VIB

GRETA-VIB is used as a virtual agent platform [Pecune et al., 2014]. GRETA-VIB follows the SAIBA framework illustrated in Figure 7.13 [Vilhjalmsson et al., 2007]. First, the “intent planner” generates the communicative intentions of the agent (performative, emotion, backchannel, etc.) and represents them in the Functional Markup Language (FML) [Heylen et al., 2008].

The “behavior planner” transforms the communicative intentions of the agent into a set of synchronized multimodal behaviors (e.g., head movements, facial expressions, gesture). The selected behaviors are represented in the Behavior Markup Language (BML) format [Vilhjalmsson et al., 2007]. Finally, the “behavior realizer” generates the animations (computed frame by frame) corresponding to the multimodal behaviors described in the BML file. For example, the “behavior realizer” computes FAPs (Facial Animation Parameters) frames to animate agent’s facial expression and BAPs (Body Animation Parameters) frame for agent’s body movement like head and torso movements. See [Pandzic and Forchheimer,] for more detail about FAPs and BAPs.

In our system, Flipper replaces the “intent planner” in the GRETA-VIB platform: Flipper generates the communicative intentions of the agent and sends the corresponding multimodal behavior in FML format to the “behavior planner”.

7.5.4 BehaviorPrediction: Agent's Behavior Prediction

We are interested in adapting agent's behavior to the user's behavior. That is, the agent not only communicates its intentions but also adapts its behavior in real time to user's behavior. Doing so the agent communicates its engagement as well. To reach this aim, we need to predict which behavior should be displayed by the agent at each moment. This is realized by adding a specific module to the system architecture. As such, the "BehaviorPrediction" module is responsible to compute the adaptive behaviors. It operates at the animation frame level, namely FAPs and BAPs.

Every time step t (i.e., every frame), "BehaviorPrediction" module computes which facial and body animation parameters (FAPs and BAPs) the agent should display at time $t+1$ (i.e., next frame). This module takes as input the user's and the agent's facial and body animation parameters as well as the conversation state over a time window ($[t-20, t[$). The learning time window is fixed to 20 frames. Prediction of the next frame is made by the underlining LSTM model. Thus, the predicted animation of the agent is computed from the user's and the agent's previous animations.

To start the prediction System, we let pass the 20 first frames of an interaction (we choose 20 frames based on the model loss as indicated in section 7.4). Then, a sliding Window determines which frames are used to predict the next frame. The first 20 frames ($f_0, ..., f_{20}$) of agent and of user are used to predict the agent's animation for the frame 21 (f_{21}). Then, the frames ($f_1, ..., f_{21}$) of the agent and of the user are used to predict the agent's animation for the frame 22 (f_{22}) and so on.

The user's behaviors are extracted using EyesWeb. EyesWeb communicates with the LSTM model by sending user's behaviors and the conversation state. To compute the conversation state (who is speaking? the agent or the user), EyesWeb relies on user's voice activity (detected with the microphone) and agent's turn (*AgentTurn*) received from Flipper.

Flipper sends the FML file to the behavior planner. The behavior planner computes the multimodal behavior of the agent and sends it to the behavior realizer that computes the animation of the agent in terms of FAPs and BAPs. Then, before sending each frame to the animation player, we merge the animation computed from the communicative intentions with the animation predicted to adapt the agent's behavior to user's behavior. We repeat this operation at every frame. This is done at the behavior realizer level with input from the "BehaviorPrediction" module. More precisely, at every time step, the "BehaviorPrediction" module receives the last 20 frames of data from EyesWeb (user's behavior and conversation state) and from the "behavior realizer". Using this data the "BehaviorPrediction" module predicts the agent's animation using our *IL-LSTM* model and sends the predicted values of FAPs and BAPs of the next frame to the "behavior realizer". This predicted animation is merged with the animation computed from the BML outputted by the Behavior Planner. At the moment, we consider only animation parameters linked to head movement (BAPs), gaze (FAPs: AU63 and AU64) and smile (FAPs: AU12) of the user and of the agent in the "IL-LSTM" module.

As indicated in Figure 7.9, the "behavior realizer" receives the agent's animation (that we call $animation_{Intentions}$) from the "behavior planner" (indicating the agent's intentions) and from the "IL-LSTM" ($animation_{Adaptation}$) (representing the adaptation of the agent's behavior to the user's behavior). Finally, the "behavior realizer" merges both animation flows ($animation_Intentions$) and ($animation_Adaptation$) in order to allow the agent to express its communicative intention and to adapt its behavior according to user's behavior.

7.6 Evaluation

In order to evaluate our generative model, we design an interactive experiment where our "BehaviorPrediction" module is used to automatically generate the agent's behavior taking into account the user's behavior. The agent, named Alice, plays the role of a virtual guide that describes an exhibition about video games to the visitors of a Sciences museums. We assume that adapting Alice's behavior according user's behavior will increase user's engagement and satisfaction. That is, our aim is to test if our model that drives the agent's behavior would enhance the interaction and user's experience.

7.6.1 Independent Variables

We test five conditions:

- *REF*: when the agent does not adapt its behavior.
- *HEAD*: when the agent adapts its head rotation according to the user's behavior.
- *SMILE*: when the agent adapts its smile according to the user's behavior.
- *GAZE*: when the agent adapts its gaze according to the user's behavior.
- *ALL*: when the agent adapts its head rotation, smile, and gaze according to the user's behavior.

The five conditions are tested in a between-subjects design.

7.6.2 Measures

Several studies focus on measuring user's engagement during human-agent interaction [van Vugt et al., 2006, Konijn and Hoorn, 2005, H. L. O'Brien and Toms, 2010, Bickmore et al., 2012]. To evaluate the engagement of the user, we rely on the I-PEFiC framework [van Vugt et al., 2006] illustrated in Figure 7.14. The I-PEFiC model initially introduced in [Konijn and Hoorn, 2005] defines three phases in the establishment of user's engagement and satisfaction during human-agent interaction: encoding, comparison, and response. In the encoding step, a virtual agent is perceived along the dimensions of ethics (good vs. bad), aesthetics (beautiful vs. ugly), epistemics (realistic vs. unrealistic) and

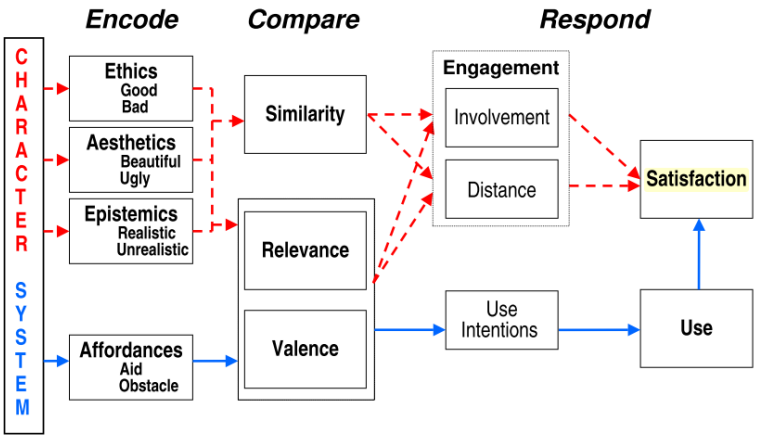


Figure 7.14 – PEFiC model [van Vugt et al., 2006]

competence (affordance). During the comparison step, user evaluates the relevance (useful vs. useless) of the agent and attributes a valence (positive vs. negative) toward the agent. Finally, during the response phase, user's engagement (i.e., involvement and distance) predicts the user's satisfaction: involvement increases the satisfaction whereas distance decreases the satisfaction (see Figure 7.14). Konijn defined involvement as felt tendency to approach the virtual agent and distance as the opposite feeling of liking [Konijn and Hoorn, 2005].

Van Vugt *et al.* proposed a questionnaire to rate the behavior of the agent according to PEFiC dimensions [van Vugt et al., 2006]. From this questionnaire, we adapt a set of items to measure the perception of the agent in term of realism (epistemics), competence and relevance (cf. table 7.4). Moreover, in order to measure the perceived friendliness of the agent, we used four items from the IAS questionnaire [Wiggins, 1979]: kind, warm, agreeable, and sympathetic. To evaluate the engagement (involvement and distance) and satisfaction of the user we rely on items described in Table 7.5). Finally, to measure the apriori attitude of participants towards the agent, we adopt eight items from the NARS questionnaire[Nomura et al., 2006]. For all items, answers are on a 5-point labeled Likert scale (1 =“strongly disagree” ... 5 = “strongly agree”).

7.6.3 Hypotheses

Previous works report that, when a robot adapts its behavior according to the user's behavior, users were more satisfied about their interaction with the robot [Fong et al., 2002, Breazeal., 2004, Woolf and Burleson, 2009]. Thus, we assume that when Alice adapts its behaviors according to the user's behavior, the user will be more satisfied about it in the interaction. Moreover as smiling is a strong cue of friendliness, we expect that the agent adapting its smile will be perceived as more friendly. Thus, our hypotheses are:

- **H.HEAD**: when Alice adapts its head rotations, the users will be **more satisfied** with the interaction compared to the users interacting with Alice in the reference condition *REF*.

Dimension	Item
Realism	Alice resembles to a real life person
Competence	Alice is effective
	Alice is qualified
	Alice is competent
	Alice is intelligent
Relevance	Alice motivated me to go and see the exhibit on video game
	Alice taught me something
Friendliness	Alice is kind
	Alice is warm
	Alice is agreeable
	Alice is sympathetic

Table 7.4 – Items used to evaluate the perception of the agent.

Dimension	Item
Involvement	Alice gives me a good feeling
Distance	I dislike Alice
Satisfaction	I am satisfied about my interaction with Alice
	I appreciated Alice
	I would like to talk again with Alice
Attitude	I would feel uneasy if virtual characters had emotions
	I would feel relaxed talking with virtual characters
	I feel comforted being with virtual characters that have emotions
	The word "virtual character" means nothing to me
	I would hate the idea that virtual characters were making judgements about things
	I would feel very nervous just standing in front of a virtual character
	I would feel paranoid talking with a virtual character
	I am concerned that virtual characters would be a bad influence on children

Table 7.5 – Items used for measuring user's engagement and satisfaction as well as user's apriori attitude toward virtual agent.

- **H.SMILE1**: when Alice adapts its smile, the users will be **more satisfied** with the interaction compared to the users interacting with Alice in *REF*.

- **H.SMILE2**: when Alice adapts its smile, it will be evaluated as **more friendly** compared to Alice in *REF*.

- **H.GAZE**: when Alice adapts its gaze, the users will be **more satisfied** with the interaction compared to the users interacting with Alice in *REF*.

- **H.ALL1**: when Alice adapts its head rotations, smile and gaze, the users will be **more satisfied** with the interaction compared to the users interacting with Alice in *REF*.

- **H.ALL2**: when Alice adapts its head rotations, smile and gaze, it will be evaluated as **more friendly** compared to Alice in *REF*.

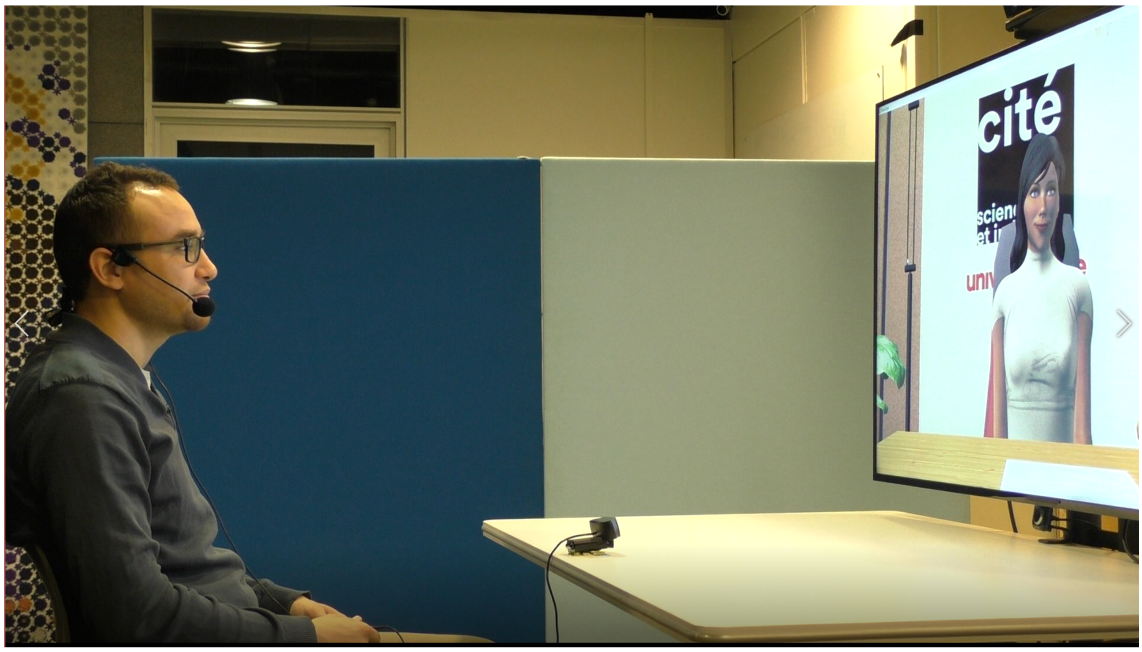


Figure 7.15 – Our system for human-agent interaction

7.6.4 Protocol

After receiving one participant in the experiment room, we first obtain his consent for using the participant's data.

Secondly, participants were asked to fill in a questionnaire about their attitudes towards virtual agents using the NARS questionnaire [Nomura et al., 2006]. Then, we gave instructions to the participant about the interaction and prepared him to interact with Alice by setting up the microphone and showing the position where to sit down.

Once the participant was ready, we started our system. At the end of the interaction, participant filled in a questionnaire to evaluate Alice as well as his satisfaction about the interaction. We also collected demographic information of the participant such as his gender, cultural identity, age and education level.

We used a Microsoft's Kinect 2 for detecting the facial expression and head movements of the user. The audio of the user is detected using a dynamic headset microphone connected through a TASCAM US-322. To increase the user's immersion, we used a large screen allowing displaying the virtual agent at a life-size scale.

7.7 Results

101 participants took part of our experiment; they are almost equally distributed among the 5 conditions. 50% are female and 95% are native French speakers. The age of participants is given in Table 7.6.

In Table 7.7, we give Cronbach's α, mean and standard deviation of each measured dimension for the five conditions.

Age group	18-25	26-35	36-45	46-55	55+
Percentage	33%	18.6%	22.7%	18.6%	7.2%

Table 7.6 – Distribution of participants w.r.t their age.

Dimension	α	REF		HEAD		SMILE		GAZE		ALL	
		M	SD	M	SD	M	SD	M	SD	M	SD
Realism	-	1.7	0.92	1.95	0.82	2.10	0.92	2.08	0.90	1.73	0.86
Competence	0.88	2.98	1.02	3.45	0.81	3.6	0.73	3.65	1.06	3.18	1.19
Relevance	0.82	2.95	1.83	3.69	0.76	3.97	0.97	3.5	1.24	3.43	1.07
Friendliness	0.89	2.98	1.13	3.01	0.86	3.80	0.83	2.90	0.86	3.86	0.81
Involvement	-	2.65	1.22	2.65	1.15	3.52	1.00	2.83	1.33	3.60	1.07
Distance	-	2.5	1.12	2.11	1.01	1.76	0.97	1.7	1.09	2.05	0.9
Satisfaction	0.81	2.95	1.38	3.75	0.76	4.1	0.79	3.5	1.1	3.82	0.91

Table 7.7 – Cronbach's α, mean and standard deviation of each measured dimension for the five conditions.

To investigate how participants rated Alice in the condition *REF* and in the other conditions, we conducted unpaired t-tests. The revealed differences are summarized in the following:

- Alice in *SMILE* condition is evaluated as more friendly ($p = .01$) than Alice in *REF* condition. Moreover, users in *SMILE* condition are more involved ($p < .01$), less distant ($p < .01$) and more satisfied about the interaction ($p = .01$) than users in *REF* condition, thus the hypotheses **H.SMILE1** and **H.SMILE2** are validated. In addition, Alice is evaluated as more positive on the relevance dimension ($p < .01$).
- Alice in *ALL* condition is evaluated as more friendly ($p < .01$) than Alice in *REF*. Also in *ALL* condition, users are more involved ($p = .01$), and less distant ($p < .01$) than users interacting with Alice in *REF*. Moreover participants in *ALL* were more satisfied with their interaction with Alice compared to the users interacting with Alice in *REF* condition, therefore **H.ALL1** and **H.ALL2** are supported.
- Alice in *HEAD* condition is evaluated as being significantly more positive on the relevance dimension ($p < .01$) compared to Alice in *REF*. No difference could be confirmed for the other dimensions including the user's satisfaction, thus the hypothesis **H.HEAD** is not accepted.
- No difference between the evaluation of Alice in *REF* and *GAZE* could be found in term of user's satisfaction. Thus, the hypothesis **H.GAZE** is rejected.

Based on the NARS questionnaire, we find that 40%, respectively 30% and 30%, of participants have a positive, respectively neutral and negative, attitude toward virtual agents. In order to investigate the effect of participant's prior attitude on agent perception, and the bias this can introduce on participant's engagement, we rely on ANOVA test. The results reveal a main effect of the participant attitude on the user's distance ($F(1, 93) = 5.13,\ p = .02$)). Bonferroni test shows that participants with negative attitude are less engaged (more distant ($p = .01$) and less involved ($p = .02$)) than those

with positive attitude. The negative attitude of participants is equally distributed between the five conditions. Consequently, the effect of attitude of participants on user's distance evaluation does not affect our results.

We designed a face-to-face interaction where the agent is displayed on large screen at a life-size scale. The user may have felt uncomfortable because of the large display of the agent. This could explain why most users stared at the agent and did not move their head and gaze. As such, the adaptive behaviors, *head movement* and *gaze*, of the agent were constant throughout the interaction. This could explain the reason our hypotheses **H.HEAD** and **H.GAZE** have been rejected.

7.8 Conclusion

When interacting with each other, we continuously analyze the behavior of our interaction partners and adapt ours accordingly. We have designed a LSTM model to adapt the agent's behavior according to the user's behavior. To integrate and evaluate our works, we have created an interaction system where the agent interacts in real-time with a human user. The system takes as input data from the user, computes what the agent has to say as well as the corresponding smile, head movements and gaze direction.

The implemented system has been evaluated using a scenario where an agent named Alice played the role of a virtual guide, presenting an exhibit about video games to museum visitors. We relied on the assumption that human users would be more satisfied by the interaction with Alice when its adapts its behavior (gaze, smile and head movement) depending on the user's behavior. The results showed that users were indeed more satisfied by their interaction with Alice when it adapted its behavior. However, these results were significant only when Alice adapted its smile to user's behavior (mainly his smile). A user-related bias could have prevented from having significant results for the other signals. During the interaction, most of the users gazed at Alice without doing any postural shift or even changing gaze and head direction. Therefore, the adaptive behaviors *head movement* and *gaze* of the agent were constant throughout the interaction. They reflected the behaviors of the user (that was not moving much).

Take home

- We developed a LSTM-based model to predict the agent's behavior in response to the user's behavior.
- We conceive a study aimed to investigate the effect of adapting agent's behavior to the user's behavior on user engagement and satisfaction.
- Participants were more satisfied about their interaction with Alice and rated Alice more positively when it adapts its behavior according to the user's behavior. However, these results are significant only when Alice smiles in response to the user.

Chapter 8

Engagement Modeling in Dyadic Interactions

Contents

8.1 Related Works **106**
8.1.1 Engagement-Related Behaviors 106
8.1.2 Engagement Prediction 106
8.2 LSTM Model for Engagement Prediction **108**
8.2.1 Data 108
8.2.2 Model 109
8.2.3 Results 112

DURING the last years, engagement modeling has gained increasing attention due to the growing number of conversational agents and the important role that engagement plays in human-agent interaction. The agent should be able to continuously detect the engagement level of the user in order to react accordingly. Our aim is to predict the engagement level of the human user during human agent interaction. Many definitions of engagement have been provided. A survey of engagement definitions in human-agent interaction is given in [Glas and Pelachaud, 2015]. Among the existing definitions, two are commonly used in human-agent interaction. One of the first definitions that was proposed is by Sidner and colleagues that define the engagement process as *"the process by which participants involved in an interaction start, maintain and terminate an interaction"* [Sidner et al., 2005]. Later on Poggi defined engagement as: *"the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction"* [Poggi, 2007]. Engagement is manifested multimodally. In particular, several works studied the role of upper face expressions such as gaze and smiling in conveying different levels of engage-

ment [Allwood and Cerrato, 2003, Castellano et al., 2009a]. But, engagement is not measured from single cues, but from the association of several cues that arise over a certain time window [Peters et al., 2005, Sidner et al., 2005, Bickmore et al., 2012]. Thus, for more accurate engagement prediction, we should take into account the sequentiality and dynamics of engagement-related signals (gaze, smile, etc.). For this we experiment LSTM networks. The LSTM takes as input a sequence of features occurring within the last n frames and predict the engagement for the next frame. The potential of LSTM is important since it can jointly model the temporality and the sequentiality.

In this Chapter, we describe our LSTM-based model for predicting user's engagement level during human agent interaction. We first present the most significant works related to engagement prediction. Then we describe the model architecture and its evaluation.

8.1 Related Works

During human-agent interaction, the user's engagement is an important aspect to be considered by the agent. For example, if the agent detects a decrease in the engagement level of the user, it should adapt its behavior in order to reengage the user in the interaction. In this section, we first, investigate what are the multimodal behaviors that participate to a change of perception of the engagement level in a human-agent interaction. Then, we present an overview of existing models of engagement prediction.

8.1.1 Engagement-Related Behaviors

Engagement can be expressed by both verbal and non-verbal behaviors. It can also be directly linked to prosodic features [Yu et al., 2004a] and verbal alignment behaviors [Pickering and Garrod, 2004]. Facial expressions are crucial indicators of engagement, for example, several studies have reported that smiling and head nod can provide information about the user's engagement level [Allwood and Cerrato, 2003, Castellano et al., 2009a, Yu et al., 2016]. Gaze is also an important cue of engagement level [Sidner et al., 2003, Peters et al., 2005, Nakano, Yukiko I. and Ishii, 2010], for example, looking at the speaking partner can be interpreted as a cue of engagement, while looking around the room may indicate the intention to disengage. Moreover, a correlation has been found between engagement and several body postures [Mota and Picard, 2003, Sanghvi et al., 2011]. Turn-taking behavior is also related to engagement as reported in [Sidner et al., 2003, Cafaro et al., 2016b].

Based on these studies, we decide to predict user's engagement based on the most popular features, namely, head movements, gaze, facial expressions (AUs). We also choose to include turn-taking (conversation state) as recommended in [Sidner et al., 2003].

8.1.2 Engagement Prediction

Over the past decade, user's engagement has been widely studied in human-agent interactions. Table 8.1 reports an overview of some works related to engagement prediction. For

each work, we indicate the technique and the features used for engagement prediction as well as the obtained results.

Reference	Goal	Feature types	Corpus	Algorithm	Class modality	Accuracy
[Forbes-Riley et al., 2012]	disengagement detection	acoustic, lexical	7K spoken dialogue turns	J48 decision trees	4	68% F-measure on *disengaged* class
[Castellano et al., 2009b]	engagement recognition with a robot companion	smile and gaze	96 samples	Bayes net.	2	94% recall on *engaged* class
[Ishii, 2010]	engagement prediction	gaze	16 minutes of human-agent interaction	3- gram patterns	2	71% F-measure
[Bohus and Horvitz, 2014]	disengagement detection	linguistic hesitation	126K frames	Logistic regression	2	89% recall on *disengaged* class
[Yu et al., 2004b]	engagement detection		7K data records	SVM, HMM	2	72% recall on *engaged class, 71% on disengaged class*
[Dhamija and Boult, 2017]	contextual engagement prediction	mood and action units	8M video frames	LSTM	5 engag. levels	55% $\pm$ 18% recall

Table 8.1 – An overview of works related to engagement prediction.

As can be seen from the table, all the works are not interested in the same goal. If they all focus on engagement modeling as a general purpose, they implement it differently depending on the task at hand, for example, an important part focus on specifically detecting disengagement. To do so, most works consider two levels of engagement (binary classification) and rely on facial expressions as predictive features. The models have been tested on different datasets and achieved very sparse results. In the binary-class case, the detection rates of engagement were 71%–95%. The heterogeneous nature of the reported experiments and results makes it hard to draw any conclusions.

Some of the reported works make use of sequential models to capture the associations among signals that may arise over a certain time window. HMMs, T-Patterns, and more recently LSTM, are examples of models that have been used for this purpose. The works of [Dhamija and Boult, 2017] remain the most similar to ours, since they use LSTM models and action units as predictive features. They also measure engagement with 5 different levels, as we do. However, they consider the detection of "contextual" and "self-reported" engagement, i.e., engagement that is constant and globally characterizes the interaction. In contrast to that, our model builds on the evolving nature of engagement during the interaction for a more fine-grained and accurate modeling. Thus, we model engagement as dynamic variable that change during an interaction as interactants became more or less

engaged with each other. Such model allows as to adapt the agent's behavior according to the predicted engagement level of the user during human-agent interaction.

8.2 LSTM Model for Engagement Prediction

In this section, we describe the developed LSTM model for engagement prediction. We detail the model architecture and evaluation. But, we first start by presenting the data and explain how it has been annotated.

8.2.1 Data

Our model has been tested on the NoXi dataset, a corpus of expert-novice interaction [Cafaro et al., 2017] (see chapter 7). We have manually annotated the engagement levels of both expert and novice, but we have semi-automatically annotated the conversation state of the interaction (who is speaking?) as detailed in Chapter 7. We relied on OpenFace framework for extracting facial expressions of both expert and novice.

Engagement annotation (manual)
We have annotated each video of the corpus with the perceived level of engagement. We have followed the recommendations described in [Yannakakis et al., 2017] to reduce and facilitate the complexity of the annotation task. Thus, five levels have been define to characterize the changes in the perception of engagement:

- Level1: strongly disengaged;
- Level2: partially disengaged;
- Level3: neutral;
- Level4: partially engaged;
- Level5: strongly engaged.

In order to avoid content biases from the verbal behavior when annotating engagement, we have filtered it out, for both expert and novice, by applying a Pass Hann Band Filter. In this way, the speech kept the prosodic information without intelligibility of its verbal content.

All 20 sessions were annotated by the same annotator except one session that has been annotated by another annotator. The agreement between the two annotators in term of Cohen's Kappa was 0.81. Table 8.2 gives the percentage of occurrence for each engagement level. As we can observe, both expert and novice are predominantly perceived by the annotators as *partially engaged* (level4).

Facial expression extraction
We have used OpenFace to extract Action units (Aus) that represent facial expression classified based on the FACS (Facial Action Coding System) taxonomy [Ekman and Friesen, 1976] (See Chapter 7 for more details about action units). We also extracted head rotation as well as the gaze of both expert and novice:

	Level1	Level2	Level3	Level4	Level5
Expert	1.25%	6.33%	14.32%	71.56%	6.51%
Novice	2.34%	10.12%	24.78%	58.05%	4.68%

Table 8.2 – Percentage of each engagement level in NoXi database for expert and novice.

- The intensity (from 0 to 5) of 17 AUs: AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU45.
- The activation (0 absent, 1 present) of 18 AUs: AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU28, AU45.
- Head rotation: rotation is in radians around X,Y,Z axes.
- Gaze direction: gaze direction in radians (gaze_angle_x, gaze_angle_y).

8.2.2 Model

We design a neural network for the purpose of predicting the engagement level of expert in NoXi database. In order to predict the engagement level, the input of our model is linked to a set of non-verbal signals: actions units, head rotation, gaze, along with the conversational state of the interaction. The model has three layers: a LSTM hidden layer to extract features from the input layer, followed by an output layer (dense) for predicting the engagement level. The Softmax function is used as activation function which is usually used for multiclass classification. For each input, softmax outputs the probability distribution of this input over the five engagement levels. For evaluation purpose, 80% of data were used for training the model, 10% for validation, and 10% for testing. A categorical cross-entropy was used to compute the loss of the model throughout the iterations. Finally, a dropout mechanism is implemented in order to prevent over-fitting. Dropout is a regularization technique for neural network models proposed in [Srivastava et al., 2014]. Dropout consists of randomly removing neurons from the neural network during training. The dropout probability was fixed to 20%.

To measure the engagement of the expert in the NoXi database we developed three LSTM models as indicated in Figure 8.1:

- Expert LSTM: this model takes as input the data originating from the expert, namely her facial expression, gaze direction, head movement and conversational state, to predict her engagement level.
- Novice LSTM: this model takes as input the data produced by the novice to predict the engagement level of the expert. The motivation behind this architecture is to explore assumptions like “whether looking at one interlocutor can be used to predict the level of engagement of the other interlocutor”.

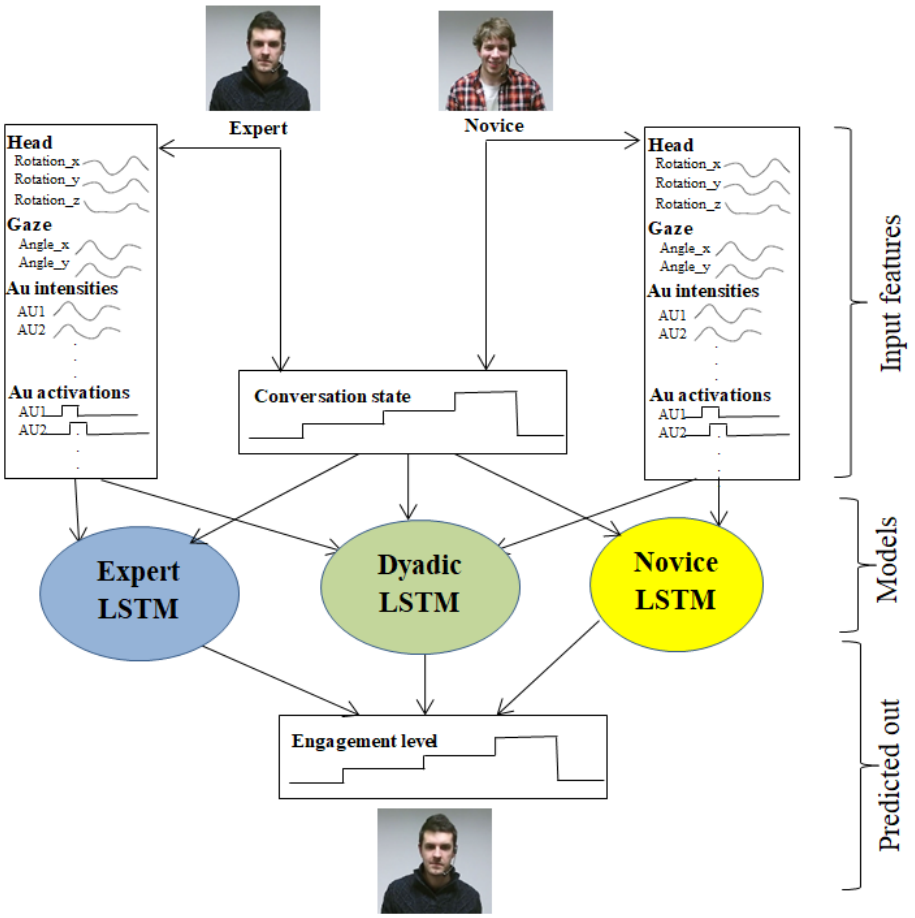


Figure 8.1 – Three LSTM models for predicting expert's engagement level.

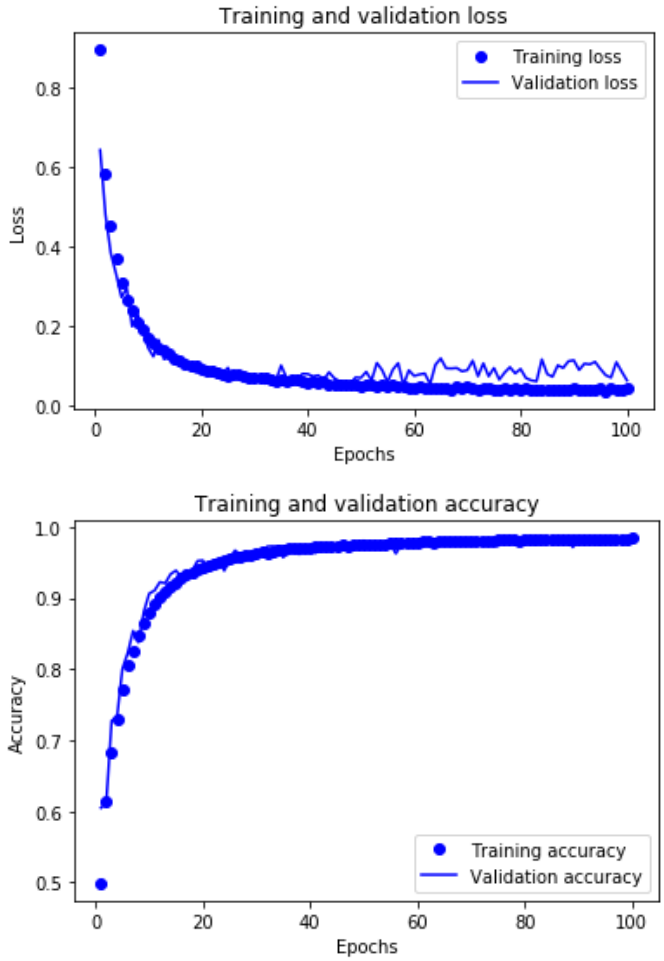


Figure 8.2 – Training and validation loss (top) and accuracy (bottom) w.r.t. the number of epochs.

- Dyadic LSTM: in this model, we merge data coming from both interaction partners, the expert and the novice, to predict the engagement level of the expert. By developing this model, we aim to validate assumptions like "whether looking at both interaction partners enhances the engagement prediction of one interlocutor". Hence, we better understand the impact of including the behaviors and the reactive behaviors of the interlocutors on the final predictions.

As can be seen from Table 8.2, our data is highly imbalanced and is mainly distributed on one major class (Level4). One popular solution to deal with this issue is either over-sampling the minority class, or under-sampling the majority class [Nitesh V. Chawla et al., 2002]. Another solution is to parameterize the LSTM (*class_weight*) with the prior distribution of classes in order to equally penalize the over and under-represented classes in the training set. We have tested these three solutions to mitigate the effect of imbalanced data. Results showed that weighting the LSTM model with prior distribution performs better than the over-sampling and under-sampling strategies.

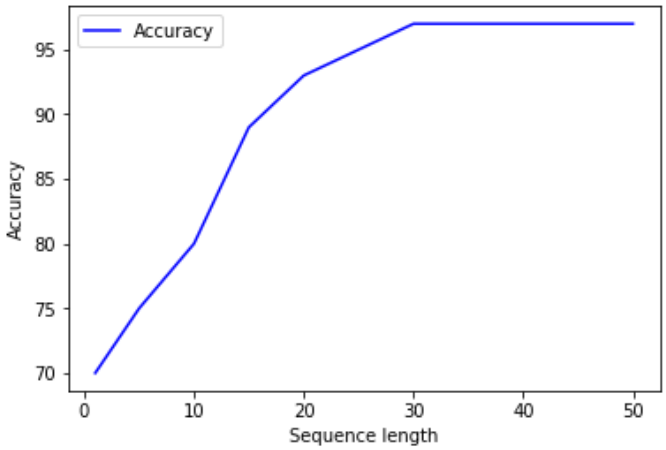


Figure 8.3 – F-measure w.r.t the sequence length given as input.

Concerning the other parameters of the model, we vary the number of epochs from 0 to 100. Figure 8.2 shows the training and validation loss for each of these values, and Figure 8.2 indicates the training and validation accuracy. Both loss and accuracy become steady after 30 epochs. Based on this observation, we train our network for 30 epochs and then evaluate it on the test set. Obtained results are given in the next section.

In LSTM model, the prediction model takes as input the last n frames of features and predicts the engagement level of the expert for the next frame. One frame corresponds to 0.04 second. To help fixing the sequence length of the input, we vary n from 0 to 50; that is we vary the number of past frames we consider to compute the engagement level at the next frame. Figure 8.3 shows the F-measure for each of these values. Based on these results, we choose 30 frames (for a total duration of 1.2 seconds) since the F-measure stops improving after this value.

8.2.3 Results

In this section, we evaluate the above-described model in terms of prediction accuracy as well as other criteria to assess the quality and relevance of the used features. Specifically, we are interested in:

- Evaluating the prediction accuracy of our model and comparing it to baseline predictors.
- Exploring the effect of dyadic features (relative to the other interaction partner) and their contribution to the global performance of engagement level prediction.
- Ranking the input features based on their relevance to the engagement prediction problem.

For the evaluation and comparison of the model, we rely on quantitative measures from Information Retrieval: recall, precision, and F-measure. For more details on calcu-

lating these measures, please refer to Chapter 5. The overall F-measure is equal to the weighted mean of each of the predicted classes.

First, we start by giving the overall accuracy values for the expert's engagement prediction using several baselines as well as our model, hence we use the Expert LSTM model. Results are given in Table 8.3. The non-mentioned parameters are left to their default values as defined in Sickit-Learn package[1].

Model	Parameters	Recall	Precision	F-measure
Random	uniform probability over classes	19%	18%	19%
Naive Bayes	-	39%	39%	37%
Random Forest	10 trees, max. tree depth=5	49%	50%	48%
AdaBoost	-	50%	51%	50%
Decision Tree	max. tree depth=5	50%	48%	47%
Neural Net	multilayer perceptron, alpha=1	68%	68%	68%
Our model	-	97%	97%	97%

Table 8.3 – Prediction of expert engagement based on different models.

From Table 8.3, we can see that our model (expert LSTM), significantly outperforms all other predictors in terms of recall, precision, and F-measure. The multilayer perceptron comes second with a difference of 30 percentage points in terms of accuracy measures. We believe that the ability of LSTM to capture the dynamic features of non-verbal behavior (namely sequentiality and temporality) is behind these good results.

To assess the contribution of each feature when taken separately (*gaze*, *head rotation*, *AU activation*, *AU intensity*, *AU*), where *AU* is the union of actions units related to either activation or intensity. We train our models by considering one feature at a time. The F-measure values obtained for each model are reported in Table 8.4. To avoid reporting too much data here, the precision and recall values are given in Appendix C.

The upper part of Table 8.4 recapitulates the obtained relevance for each feature type. We observe that the feature *AU* (i.e. AU activation and intensity) is the most informative for this task. It achieves single-handedly 91%, whereas *head rotation* and *gaze* achieve only 63% and 62%, respectively. Moreover, relying on Action Units the model is able to predict all engagement levels with very high accuracy. The features *head rotation* or *gaze*, when considered alone, predict very poorly 4 engagement levels out of 5. They often predict the engagement level 4 that represents the majority class. The features *AU intensity* and *AU activation* are approximately equally discriminant for engagement prediction when taken separately. Their combination improves the overall accuracy by 15 percentage points.

On the other hand, by considering only the novice's behavior (while ignoring the expert's behavior, as shown in the middle pat of Table 8.4), the model achieves a fairly good result, almost equivalent to the configuration considering only the expert's behavior. This result is consistent with other findings from [Nguyen et al., 2014] where the best predictors of job interview success for an applicant were those related to the other partner's (recruiter's) behavior. Finally, considering the behaviors of both, novice and expert,

[1]https://scikit-learn.org/

Features	Weighted F-measure	F-measure specific to each engagement level				
		Level1	Level2	Level3	Level4	Level5
Expert LSTM						
Gaze	62%	0%	31%	7%	83%	0%
Head	63%	0%	35%	10%	84%	0%
AU intensity	76%	42%	48%	49%	88%	41%
AU activation	73%	45%	46%	48%	86%	22%
AU	91%	86%	86%	84%	95%	75%
All features	97%	96%	95%	94%	98%	93%
Novice LSTM						
Gaze	60%	0%	11%	9%	83%	0%
Head	61%	0	0.17	0.08	83%	0%
AU intensity	77%	7%	57%	53%	89%	45%
AU activation	71%	1%	49%	38%	86%	29%
AU	92%	79%	85%	84%	95%	79%
All features	96%	93%	94%	94%	98%	91%
Dyadic LSTM						
Gaze	72%	6%	46%	38%	86%	40%
Head	72%	2%	47%	44%	87%	15%
AU intensity	94%	88%	85%	87%	97%	88%
AU activation	92%	86%	87%	85%	95%	80%
AU	98%	96%	95%	95%	98%	96%
All features	99%	95%	96%	96%	99%	97%

Table 8.4 – Prediction of expert's engagement level using each feature separately (in addition to their union) for the three different configurations.

simultaneously (see Table 8.4, bottom part) significantly improves the recognition rates. We can note that the prediction based solely on the feature AU gives an equivalent result compared to the configuration involving all features.

In this chapter, we focused on an important aspect of human-agent interaction: the engagement that ensures the interaction to go on. After investigating which behaviors participate to a change in engagement perception, we focused on facial expressions, head movements, and turn taking that represent a relevant set of indicators to engagement prediction. We developed an LSTM model to predict the engagement level of the user. We trained our model using the NoXi database involving an expert and a novice engaged in a conversation. The engagement levels of both have been manually annotated. We explored the contribution of several multimodal features of the expert, namely expert's gaze, head and action units, and their discriminating power for the task of engagement prediction. We also investigated the importance of considering the novice's behavior for predicting the engagement of the expert. The obtained results revealed that Action Units contribute more than gaze and head movement to the prediction of engagement. Moreover, the results confirmed the importance of considering both partners' behavior in a dyadic interaction for engagement prediction. Our model (Expert LSTM) is used in an ECA platform to

continuously feed the ECA with the predicted levels of engagement of the user[Mancini et al., 2019].

Take home

- Action Units contribute more than gaze and head movement to the prediction of engagement.
- Considering only novice's behavior, while ignoring expert's behavior gives a fairly good result to predict engagement, almost equivalent to the configuration of considering only the expert's behavior.
- Considering both novice's and expert's behaviors significantly improves the recognition rate.
- These results confirm the importance of considering both partners' behaviors in a dyadic interaction for engagement prediction.
- Our prediction model is integrated in human-agent interaction in order to predict, in real time, the engagement level of the user.

Chapter 9

Conclusion

Contents

9.1 Summary **117**
9.1.1 Attitude Variation Modeling 117
9.1.2 Adapting Agent's Behavior According to the User's Behavior . . 119
9.1.3 Engagement Prediction 120
9.2 Limits and Perspectives **120**

A social interaction implies a social exchange between two or more persons, where they adapt and adjust their behaviors in response to their interaction partners. With the growing interest in human-agent interactions, it is desirable to make these interactions natural and human like. In this context, we aimed at enhancing the quality of the interaction between users and ECAs by endowing an ECA with the capacity to (1) express different social attitudes, (2) adapt in real time its behavior according the user's behavior and its communicative intentions, and (3) predict, in real time, the engagement level of the user. To achieve all these goals we have leveraged the *sequentiality* and *temporality* of non-verbal behaviors and relied on appropriate techniques: *temporal sequence mining* and *recurrent neural networks*.

9.1 Summary

9.1.1 Attitude Variation Modeling

Within a human-agent interaction, the agent should be able to adapt and vary its attitude toward the user according to its role, the behavior of the user, etc. Number of attempts have been made to model the attitude of a virtual agent based on its non-verbal behavior. However, these models ignore the temporal information, e.g., starting time and duration

of non-verbal signals, which could influence and change the perception of these behaviors as discussed in Chapter 2.5. The novelty of our work was to consider *temporality* of non-verbal signals for attitude modeling. It is recalled that our aim was to extract the most relevant sequences of non-verbal behaviors that trigger a variation in attitude perception (e.g., what makes the agent appear more friendly?). Thus, we have represented an attitude variation as a temporal sequence of non-verbal signals. We have naturally opted for sequence mining to extract the most relevant temporal sequences representing attitude variations.

Temporal sequence mining algorithms build on a predefined temporal distance between events then group similar events in the same cluster accordingly. The decision to group similar events is made when the distance comes in under a certain threshold. Existing temporal sequence mining algorithms have made the choice to fix this threshold once and for all. Such a choice may turn inefficient, for example, when dealing with real data. In the case of non-verbal behavior, they do not consider the intrinsic differences between non-verbal signals that are of the same type. For example, the duration of head movements are typically much shorter than the duration of postures. We have addressed this limitation by introducing a new temporal sequence mining algorithm, named HCApriori, specifically designed to overcome the shortcomings of existing algorithms.

HCApriori considers the differences between event types and consequently increases cluster homogeneity. By comparing HCApriori to the state-of-the-art, we found that our algorithm significantly overcomes the existing algorithms in terms of extraction accuracy, defined as the percentage of sequences from the original data that are similar to at least one pattern from the set of extracted patterns. At the same time, we have improved the standard metrics, *support* and *confidence*, that reflect the quality of the extracted patterns. Note that these metrics are originally only based on the occurrence frequency of events. We have extended them by incorporating the temporality criterion for a more relevant and fairer evaluation.

Building on HCApriori, we have designed a model to extract the temporal sequences of non-verbal signals conveying four attitude variations in human data: dominance decrease, dominance increase, friendliness decrease, and friendliness increase. The extracted temporal sequences have been simulated in a virtual agent and evaluated through a perceptive study. In contrast to previous works, we have evaluated both dimensions of attitudes at the same time (dominance and friendliness), which allowed studying the interrelation between the perception of attitudes. By doing that, we have discovered a compensation effect between the perception of dominance and of friendliness in a virtual agent. We have also found a high correlation between dominance increase and friendliness decrease, which could be explained by the fact that both variation types result from the same non-verbal behaviors. On the other hand, we were surprised to find out that the patterns conveying “no attitude” when played by the virtual agent were perceived as expressing friendliness. The main take-home message from this study was that the extracted patterns for the different attitude variations from human data, when applied to virtual agents, were mostly perceived as such by humans.

The integration of this attitude behavior model has been realized through an *attitude planner* to allow virtual agents expressing different attitude variations. This planner takes as input the attitude variation the agent should express as well as its communicative intentions. First it converts the communicative intentions into a sequence of non-verbal behaviors, then selects the most relevant patterns that represent the desired attitude variation. Then, the two sequences are merged into one final sequence conveying both the communicative intentions and the attitude variation of the agent.

We have used the developed attitude planner to generate the behavior of an ECA taking as input different attitude variations. Then, a perceptive study was conducted to assess whether the generated behavior is appropriately recognized as truly expressing an attitude variation. The results showed that thanks to our attitude planner, the ECA became able to appropriately express a number of attitude variations, in particular *dominance increase* and *friendliness decrease*. The recognition fails when it comes to the *friendliness increase* variation. The high correlation between friendliness perception versus "no attitude" seems to affect the recognition of this variation. The non recognition of the variation *dominance decrease* could be caused by the role played by the agent (here a job recruiter). Acting as a virtual recruiter could have inferred to the agent a prevalent dominant attitude.

9.1.2 Adapting Agent's Behavior According to the User's Behavior

When interacting with each other, we continuously analyze the behavior of our interaction partners and adapt ours accordingly. In the context of human-agent interaction, most existing works infer the affective state of the user (e.g., joy, anger, surprise, fear, disgust, sadness) and adjust the facial expression of the agent appropriately. Feng *et al.* proposed another approach to directly predict the facial expressions of one partner in response to the facial expression of the other partner in dyadic interaction [Feng et al., 2017]. Thus, the facial expression of one partner is predicted from both partners' facial expression. This model was not used to predict the behavior of virtual agent in real time.

We have designed a LSTM model that we called IL-LSTM (Interaction Loop LSTM) to adapt the agent's facial expression according to the user's facial expression. The novelty of our model was the prediction of agent's facial expression as a function of both agent's and user's facial expression. To integrate and evaluate our works, we have created an interaction system where the agent interacts in real-time with a human user. The system takes as input data from the user, computes what the agent has to say as well as the corresponding animation. It is built upon four modules:

- User's behavior detection and analysis based on the multimodal analysis software EyesWeb.
- Dialogue management: we have used the dialog manager Flipper to define the dialogue rules (turn taking and verbal content) for the virtual agent.
- Agent's behavior prediction based on IL-LSTM to predict the behavior of the agent for the next frame taking as input the behavior of both agent and user, during the past frames.

- Behaviour generation from the predicted agent's behavior using the GRETA-VIB platform.

The implemented system has been evaluated using a scenario where an agent named Alice played the role of a virtual guide, and presented an exhibit about video games to museum visitors. We relied on the assumption that human users would be more satisfied by the interaction with Alice when its adapts its behavior (gaze, smile and head movement) depending on the user's behavior. The results showed that users were indeed more satisfied by their interaction with Alice when it adapted its behavior. However, these results were significant only when Alice adapted its smile to user's behavior. A user-related bias could have prevented from having significant results for the other expressions. During the interaction, most of the users gazed at Alice without doing any postural shift or even changing gaze and head direction. Therefore, the adaptive behaviors *head movement* and *gaze* of the agent were relatively constant throughout the interaction. They reflected the behaviors of the user.

9.1.3 Engagement Prediction

In this part of the thesis, we have shed light on an important aspect of human-agent interaction: *engagement*. Engagement ensures the interaction to go on without loss of interest or motivation. After investigating which behaviors contribute the most to a change of engagement perception, we have focused on facial expressions, head movements, and turn taking. Those three features represent relevant indicators of engagement. We have developed a specific LSTM-based model to predict the engagement level of the user, and trained it on the NoXi database containing expert-novice conversations. We have explored the contribution of different multimodal features, namely gaze, head and action units, to the engagement prediction. Results revealed that action units contribute more than gaze and head movement to the prediction of engagement. We have also investigated the importance of considering one interlocutor's behavior for predicting the engagement of the interlocutor. The results underlined the importance of considering both partners' behavior in a dyadic interaction for engagement prediction. Our model has been integrated in an ECA platform, which allowed us to continuously feed the ECA with real-time predictions of the user engagement level.

9.2 Limits and Perspectives

In the first part of this work, we have modeled attitude variations of ECA while holding the speaking turn, but not when being the listener. The same methodology can be used to design an agent conveying attitude variations when listening to the other interaction partner. After building the sequences representing attitude variations when a person is the addressee, we could use HCApriori to extract the most relevant patterns expressing these attitude variations. Regarding the attitude planner, we need to add a new input to indicate

the turn of the agent (speaking or listening) in order to select the pattern the agent will display according to both its attitude variation and its role in the turn taking.

Attitudes are expressed both through verbal and non-verbal behaviors. Several works showed that combining both non-verbal and verbal modalities led to better attitude recognition of ECAs [Bee et al., , Chollet et al., 2017]. Our model has only focused on the non-verbal behavior for attitude expression. In [Callejas et al., 2014], the authors proposed a model to express attitudes verbally. Aspects such as the length of sentences, the variety of vocabulary or the quantity of pronouns can be taken into account in order to characterize the perceived attitude of a sentence. For example, a sentence expressed with a hostile attitude may be longer, may provide more details, and emphasize negative content (e.g., using negative expression such as "not earlier than"). One can extend this type of verbal models, such as [Callejas et al., 2014], and combine it with ours in order to allow ECA to express attitude variations through both verbal and non verbal behaviors.

In our work, we have focused on attitude variation modeling, i.e., an increase or a decrease of an attitude. These variations have a given intensity (small, large, etc.). Thus, the perception of attitude variations could be influenced by its intensity. In our work, we did not consider the intensity of variation. In the future, we intend to investigate how the perception of an attitude variation is influenced by the intensity of this variation, then extend our attitude planner to allow the agent to express attitude variations with different intensity levels.

We found a high correlation between dominance increase and friendliness decrease. Both attitude variations were perceived as conveying dominance and hostility. An explanation is that some non-verbal signals have the same effect on the perception of dominance and of hostility [Knutson, 1996, Tiedens et al., 2000, Carney et al., 2005, Ravenet et al., 2013]. This assumption needs to be more thoroughly analyzed and validated.

To our surprise, the extracted patterns expressing "no attitude" (attitude value around zero) were evaluated as conveying friendliness. This could be caused by the annotation scheme that has been used to annotate the perception of attitudes. Annotators have continuously indicated the values (between 0 and 1) of the perceived attitudes. As reported by Yannakakis [Yannakakis, 2018], a drawback with continuous annotation is to provide low degree of reliability between annotators. To address this issue, the same authors proposed AffectRank: a rank-based annotation tool. This annotation approach should yield significantly less noise and higher inter-annotator agreement [Yannakakis and Martinez, 2015, Yannakakis et al., 2017]. We plan to use AffectRank to better annotate the perceived attitudes on the circumplex as indicated in Figure 9.1. For example, the *LM (friendliness)* octant could be annotated with "no" friendliness, small friendliness, and large friendliness.

The new annotation scheme, combined with our attitude modeling methodology, would improve our model by:

- Annotating both attitude dimensions at once when selecting the perceived attitude on the circumplex.

- Extracting more accurate sequences representing each attitude octant (*PA, BC*, etc.).

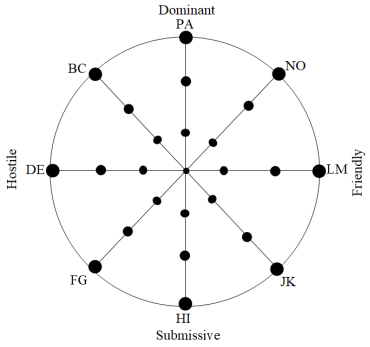


Figure 9.1 – An example of discrete annotation of attitude using the IPC with three levels for each octant.

- Better modeling of attitude intensity: for the moment, we do not consider the intensity (large, small, etc.) of attitude variations. Using the new annotation scheme, we can extract sequences relative to the different intensities of attitude variations.

Concerning behavior adaptation from the gaze and head movement, a bias has been observed. When the agent adapted its gaze or its head movements, it was perceived in the same way as if it did not. This result could be caused by the adopted evaluation scenario where agent and user gazed at each other during the whole interaction. In the future, we plan to change and enrich the interaction scenarios (e.g., collaborative task) where both agent and user will gaze at different objects. In such a setting, we expect the participants will also perform much more head movements.

We designed a model to generate the agent's facial expressions according to the user's facial expressions. Our model consider only smile, gaze and head movement. In the next future, we plan to consider more facial expressions such as eyebrow movements.

Our model allows the agent to adapt its facial expression but not its speech. We plan to enhance our interactive system by adapting the agent's speech according to the user's speech. We also plan to enhance our interactive system by adapting the agent's speech according to the user's speech using alignment model such as those proposed in [Campano et al., 2015].

Our engagement model is integrated in human-agent interaction in order to predict, in real time, the engagement level of the user. This model relies solely on the user's behavior for predicting her engagement level. However, the results obtained from another experiment confirmed the importance of considering both partners' behaviors in a dyadic interaction (user and agent) for engagement prediction. Hence the importance of considering both user's and agent's behaviors for user's engagement prediction.

Engagement can be predicted from other behaviors like prosody [Yu et al., 2004a] as well as gestures [Sidner et al., 2003]. We only investigated the relevance of gaze, action units and head movement features for engagement level prediction. As future work, we plan to extend this investigation by considering others features like gesture and prosody.

Bibliography

[Abele, 1986] Abele, A. (1986). Functions of gaze in social interaction: Communication and monitoring. *Journal of Nonverbal Behavior*, 10(2):83–101.

[Acton and Revelle, 2014] Acton, G. S. and Revelle, W. (2014). Interpersonal Personality Measures Show Circumplex Interpersonal Personality Measures Show Circumplex Structure Based on New Psychometric Criteria.

[Alden et al., 1990] Alden, L. E., Wiggins, J. S., and Pincus, A. L. (1990). Construction of Circumplex Scales for the Inventory of Interpersonal Problems. *Journal of Personality Assessment*, 55(3-4):521–536.

[Allwood and Cerrato, 2003] Allwood, J. and Cerrato, L. (2003). A study of gestural feedback expressions. In *First Nordic Symposium on Multimodal Communication*, pages 7–22.

[Arai and Hasegawa, 2004] Arai, F. and Hasegawa, Y. (2004). Facial Expressive Robotic Head System for Human – Robot Communication and Its. (December).

[Argyle, 1988] Argyle, M. (1988). *Bodily Communication*. Methuen.

[Argyle, M., Dean, 1965] Argyle, M., Dean, J. (1965). Eye contact, distance and affiliation. *Sociometry,*, (28):289–304.

[Ballin et al., 2004] Ballin, D., Gillies, M. F., and Crabtree, I. B. (2004). A Framework for Interpersonal Attitude and Non-Verbal Communication in Improvisational Visual Media Production. In *in Proceedings of the 1st European Confer- ence on Visual Media Production*, pages 203–210.

[Baltrusaitis et al., 2016] Baltrusaitis, T., Robinson, P., and Morency, L. P. (2016). OpenFace: An open source facial behavior analysis toolkit. *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*.

[Barbulescu et al., 2015] Barbulescu, A., Barbulescu, A., Generation, A., and Barbulescu, A. (2015). Audiovisual Generation of Social Attitudes from Neutral Stimuli. (SEPTEMBER).

[Beattie, 1981] Beattie, G. W. (1981). Interruption in conversational interaction, and its relation to the sex and status of the interactants. *Linguistics*, 19(1-2):15–36.

[Bee et al., 2009] Bee, N., Franke, S., and Andŕe, E. (2009). Relations between facial display, eye gaze and head tilt: Dominance perception variations of virtual agents. *Proceedings - 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009*.

[Bee et al.,] Bee, N., Pollock, C., André, E., and Walker, M. Bossy or Wimpy: Expressing Social Dominance by Combining Gaze and Linguistic Behaviors. In *Proceedings of Intelligent Virtual Agents: 10th International Conference, (IVA 2010)*, pages 265–271, Philadelphia, PA, USA.

[Bengio, 2012] Bengio, Y. (2012). Practical Recommendations for Gradient-Based Training of Deep Architectures.

[Bengio et al., 1994] Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.

[Biber, 2006] Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes*, 5(2):97–116.

[Bickmore et al., 2012] Bickmore, T., Schulman, D., and Yin, L. (2012). Maintaining Engagement in Long-term Interventions with Relational Agents. *International Society of Differentiation*, 83(2):1–29.

[Bohus and Horvitz, 2014] Bohus, D. and Horvitz, E. (2014). Managing Human-Robot Engagement with Forecasts and... um ... Hesitations. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 2–9.

[Bradley, 1958] Bradley, J. V. (1958). Complete counterbalancing of immediate sequential e ects in a latin square design. *American Statistical Association*, 53(282):525–528.

[Breazeal., 2004] Breazeal., C. (2004). Function meets style: insights from emotion theory applied to HRI. *IEEE Transactions on Systems, Man, and Cybernetics*, 34(2).

[Breunig et al., 2000] Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. In *2000 ACM SIGMOD International Conference on Management of Data (SIGMOD '00)*, pages 93–104.

[Burgoon et al., 1984] Burgoon, J. K., Buller, D. B., Hale, J. L., and Turck, M. A. (1984). Relational Messages Associated With Nonverbal Behaviors. *Human Communication Research*, 10(3):351–378.

[Burgoon and Hale, 1984] Burgoon, J. K. and Hale, J. L. (1984). The fundamental topoi of relational messages. In *Communication Monographs*, pages 193–214.

[Burgoon et al., 2010] Burgoon, J. K., Stern, L. A., and Dillman, L. (2010). Adaptation in Dyadic Interaction: Defining and Operationalizing Patterns of Reciprocity and Compensation. *Communication Theory*, (1993).

[Burgoon, J. K. and Le Poire, 1999] Burgoon, J. K. and Le Poire, B. A. (1999). Nonverbal cues and interpersonal judgments : Participant and observer perception of intimacy , dominance , composure and formality. *Communication Monographs*, 15(3):105–124.

[Cafaro et al., 2016a] Cafaro, A., Glas, N., and Pelachaud, C. (2016a). The Effects of Interrupting Behavior on Interpersonal Attitude and Engagement in Dyadic Interactions. *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems*, pages 911–920.

[Cafaro et al., 2016b] Cafaro, A., Ravenet, B., Ochs, M., Vilhjálmsson, H. H., and Pelachaud, C. (2016b). The Effects of Interpersonal Attitude of a Group of Agents on User's Presence and Proxemics Behavior. *ACM Transactions on Interactive Intelligent Systems*, 6(2):1–33.

[Cafaro et al., 2012] Cafaro, A., Vilhjálmsson, H. H., Bickmore, T., Heylen, D., Jóhannsdóttir, K. R., and Valgarosson, G. S. (2012). First impressions: Users' judgments of virtual agents' personality and interpersonal attitude in first encounters. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7502 LNAI:67–80.

[Cafaro et al., 2017] Cafaro, A., Wagner, J., Baur, T., Dermouche, S., Torres, M. T., Pelachaud, C., Andr, E., and Valstar, M. (2017). The NoXi Database : Multimodal Recordings of Mediated Novice-Expert Interactions. In *ICMI'17,*, pages 350–359, Glasgow, Scotland. ACM.

[Callejas et al., 2014] Callejas, Z., Ravenet, B., Ochs, M., and Pelachaud, C. (2014). A computational model of social attitudes for a virtual recruiter. *13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014*, 1:93–100.

[Campano et al., 2015] Campano, S., Langlet, C., Glas, N., and Pelachaud, C. (2015). An ECA Expressing Appreciations. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 962–967.

[Carney et al., 2005] Carney, D. R., Hall, J. A., and LeBeau, L. S. (2005). Beliefs about the nonverbal expression of social power. *Journal of Nonverbal Behavior*, 29(2):105–123.

[Castellano et al., 2009a] Castellano, G., Pereira, A., Leite, I., Paiva, A., and McOwan, P. W. (2009a). Detecting user engagement with a robot companion using task and social interaction-based features. *Proceedings of the 2009 international conference on Multimodal interfaces - ICMI-MLMI '09*, (January 2009):119.

[Castellano et al., 2009b] Castellano, G., Pereira, A., Leite, I., Paiva, A., and McOwan, P. W. (2009b). Detecting user engagement with a robot companion using task and social interaction-based features. (January):119.

[Castellano et al., 2009c] Castellano, G., Pereira, A., Leite, I., Paiva, A., and Mcowan, P. W. (2009c). Detecting User Engagement with a Robot Companion Using Task and Social Interaction-based Features Interaction scenario. *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 119–125.

[Chen et al., 2003] Chen, Y. L., Chiang, M. C., and Ko, M. T. (2003). Discovering time-interval sequential patterns in sequence databases. *Expert Systems with Applications*, 25(3):343–354.

[Chindamo et al., 2009] Chindamo, M., Allwood, J., and Ahlsén, E. (2009). Some Suggestions for the Study of Stance in Communication. *2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 617–622.

[Chiu et al., 2015] Chiu, C.-C., , Morency, L.-P., , and Marsella, S. (2015). No TitlePredicting Co-verbal Gestures: A Deep and Temporal Modeling Approac. In Brinkman, W.-P., , Broekens, J., , and Heylen, D., editors, *Intelligent Virtual Agents*, pages 152–166. Springer International Publishing.

[Chollet, 2017] Chollet, F. (2017). *Deep Learning with Python*.

[Chollet, 2015] Chollet, M. (2015). THÈSE TELECOM ParisTech Agents Conversationnels Animés pour l ' entrainement social : modèle computationnel de l ' expression d ' attitudes sociales par des séquences de signaux non-verbaux.

[Chollet et al., 2014a] Chollet, M., Ochs, M., and Pelachaud, C. (2014a). From Non-verbal Signals Sequence Mining to Bayesian Networks for Interpersonal Attitudes Expression. In *14th International Conference on Intelligent Virtual Agents (IVA 2014)*, pages 120–133.

[Chollet et al., 2014b] Chollet, M., Ochs, M., and Pelachaud, C. (2014b). Mining a multimodal corpus for non-verbal behavior sequences conveying attitudes. In *9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3417–3424.

[Chollet et al., 2017] Chollet, M., Ochs, M., and Pelachaud, C. (2017). A Methodology for the Automatic Extraction and Generation of Non-Verbal Signals Sequences Conveying Interpersonal Attitudes. *IEEE Transactions on Affective Computing*, (September):1–1.

[Clark, T. L., & Taulbee, 1981] Clark, T. L., & Taulbee, E. S. (1981). A comprehensive and indexed bibliography of the Interpersonal Check List. *Journal of Personality Assessment*, (45):505–525.

[Cosnier, 1997] Cosnier, J. (1997). Sémiotique des gestes communicatifs. *Nouveaux actes sémiotiques*, 52:7–28.

[Cowie et al., 2012] Cowie, R., McKeown, G., and Douglas-Cowie, E. (2012). Tracing Emotion: An Overview. *International Journal of Synthetic Emotions*, 3(1):1–17.

[Debras and Cienki, 2012] Debras, C. and Cienki, A. (2012). Some uses of head tilts and shoulder shrugs during human interaction, and their relation to stancetaking. *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, (August):932–937.

[Dermouche, 2016] Dermouche, S. (2016). Computational Model for Interpersonal Attitude Expression. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI 2016)*, pages 554–558, Tokyo, Japan. ACM.

[Dermouche et al., 2018] Dermouche, S., Curie, M., Pelachaud, C., and Curie, M. (2018). Extraction and Generation of Non-Verbal Signals Sequences Expressing Social Attitudes. In *WACAI*, Porquerolles, France.

[Dermouche and Pelachaud, 2016a] Dermouche, S. and Pelachaud, C. (2016a). Multimodal behavior modeling for social agents. In *WACAI*, Brest, France. ACM.

[Dermouche and Pelachaud, 2016b] Dermouche, S. and Pelachaud, C. (2016b). Sequence-Based Multimodal Behavior Modeling for Social Agents. In *roceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI 2016)*, Tokyo, Japan.

[Dermouche and Pelachaud, 2018a] Dermouche, S. and Pelachaud, C. (2018a). Attitude Modeling for Virtual Character Based on Temporal Sequence Mining: Extraction and Evaluation. In *Proceedings of the 5th International Conference on Movement and Computing*, pages 1–8, Genoa, Italy. ACM.

[Dermouche and Pelachaud, 2018b] Dermouche, S. and Pelachaud, C. (2018b). Expert-Novice Interaction : Annotation and Analysis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. ELRA.

[Dermouche and Pelachaud, 2018c] Dermouche, S. and Pelachaud, C. (2018c). From Analysis to Modeling of Engagement as Sequences of Multimodal Behaviors. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*, pages 786–791, Miyazaki, Japan. ELRA.

[Dhamija and Boult, 2017] Dhamija, S. and Boult, T. E. (2017). Automated Mood-aware Engagement Prediction.

[Du Bois, 2007] Du Bois, J. W. (2007). The stance triangle. *Stancetaking in discourse: Subjectivity, evaluation, interaction*, 164(3):139–182.

[Duncan, St. jr., Fiske, 1977] Duncan, St. jr., Fiske, D. (1977). Face-to-Face interaction: Research, methods and theory. *Hillsdale, N.J.: Lawrence Erlbaum Ass*.

[Ekman and Friesen, 1969] Ekman, P. and Friesen, W. (1969). The repertoire of nonverbal behavior: Categories, origins, usage and coding. *Semiotica*, 1(1):49–98.

[Ekman and Friesen, 1976] Ekman, P. and Friesen, W. V. (1976). Mesauring facial movement.pdf.

[Ekman and Friesen, 1982] Ekman, P. and Friesen, W. V. (1982). Felt , False , and Miserable Smiles. *Journal of Nonverbal Behavior*, 6(4):238–252.

[Feng et al., 2017] Feng, W., Kannan, A., Gkioxari, G., and Zitnick, C. L. (2017). Learn2Smile: Learning non-verbal interaction through observation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, volume 2017-Septe, pages 4131–4138.

[Fong et al., 2002] Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2002). A Survey of Socially Interactive Robots : Concepts , Design , and Applications Terrence Fong , Illah Nourbakhsh , and Kerstin Dautenhahn. *Technical Report CMU-RI-TR-02-29, Robotics Institute, Pittsburgh, PA*, (November).

[Forbes-Riley et al., 2012] Forbes-Riley, K., Litman, D., Friedberg, H., and Drummond, J. (2012). Intrinsic and extrinsic evaluation of an automatic user disengagement detector for an uncertainty-adaptive spoken dialogue system. *Proc. NAACL-HLT*, pages 91–102.

[Fricker et al., 2011] Fricker, D., Zhang, H., and Yu, C. (2011). Sequential pattern mining of multimodal data streams in dyadic interactions. In *IEEE International Conference on Development and Learning (ICDL 2011)*.

[Gifford, 1991] Gifford, R. (1991). Mapping nonverbal behavior on the interpersonal circle. *Journal of Personality and Social Psychology*, 61(2):279–288.

[Gifford and Hine, 1994] Gifford, R. and Hine, D. W. (1994). The role of verbal behavior in the encoding and decoding of interpersonal dispositions.

[Glas and Pelachaud, 2015] Glas, N. and Pelachaud, C. (2015). Definitions of engagement in human-agent interaction. In *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, pages 944–949.

[Goodfellow et al., 2014] Goodfellow, I. J., Pouget-abadie, J., Mirza, M., Xu, B., Wardefarley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. *In Advances in Neural Information Processing Systems*.

[Gordon et al., 2016] Gordon, G., Spaulding, S., Westlund, K., Lee, J., Plummer, L., Martinez, M., and Das, M. (2016). Affective Personalization of a Social Robot Tutor for Children ' s Second Language Skills. (2011):3951–3957.

[Guillame-Bert and Crowley, 2012] Guillame-Bert, M. and Crowley, J. L. (2012). Learning Temporal Association Rules on Symbolic Time Sequences. *Proceedings of the 4th Asian Conference on Machine Learning*, 25:159–174.

[Gurtman, 2009a] Gurtman, M. B. (2009a). Exploring Personality with the Interpersonal Circumplex. *Social and Personality Psychology Compass*, 3(4):601–619.

[Gurtman, 2009b] Gurtman, M. B. (2009b). Exploring Personality with the Interpersonal Circumplex. *Social and Personality Psychology Compass*, 3(4):601–619.

[Gurtman and Balakrishnan, 1998] Gurtman, M. B. and Balakrishnan, J. D. (1998). Circular measurement redux: The analysis and interpretation of interpersonal circle profiles. *Clinical Psychology: Science and Practice*, 5(3):344–360.

[Guyet and Quiniou, 2008] Guyet, T. and Quiniou, R. (2008). Mining temporal patterns with quantitative intervals. In *IEEE International Conference on Data Mining Workshops (ICDM Workshops 2008)*, pages 218–227.

[Guyet and Quiniou, 2011] Guyet, T. and Quiniou, R. (2011). Extracting temporal patterns from interval-based sequences. In *International Joint Conference on Artificial Intelligence (IJCAI 2011)*, pages 1306–1311.

[H. L. O'Brien and Toms, 2010] H. L. O'Brien and Toms, E. G. (2010). What is User Engagement? A Conceptual Framework for Defining User Engagement with Technology Heather. *Journal of the American Society for Information Science*, 1(6):2581–2583.

[Hall et al., 2005] Hall, J. A., Coats, E. J., and LeBeau, L. S. (2005). Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological Bulletin*, 131(6):898–924.

[Heise, 2010] Heise, D. R. (2010). *Surveying cultures: Discovering shared conceptions and sentiments.* John Wiley & Sons, 23 edition.

[Heylen et al., 2007] Heylen, D., Bevacqua, E., Tellier, M., and Pelachaud, C. (2007). Searching for Prototypical Facial Feedback Signals. In *Intelligent Virtual Agents*, pages 147–153.

[Heylen et al., 2008] Heylen, D., Kopp, S., Marsella, S. C., Pelachaud, C., and Vilhjálmsson, H. (2008). The next step towards a function markup language. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5208 LNAI:270–280.

[Hill et al., 2013] Hill, C., Adams, B. G., Meiring, D., Van de Vijver, F. J., De Bruin, G. P., Nel, J. A., and Valchev, V. H. (2013). Developing and testing items for the South African Personality Inventory (SAPI). *SA Journal of Industrial Psychology*, 39(1):1–13.

[Hochreiter and Urgen Schmidhuber, 1997] Hochreiter, S. and Urgen Schmidhuber, J. (1997). Ltsm. *Neural Computation*, 9(8):1735–1780.

[Höppner, 2002] Höppner, F. (2002). Learning Dependencies in Multivariate Time Series. *Proc., Workshop on Knowledge Discovery in (Spatio-) Temporal Data*, pages 25–31.

[Horowitz, 1997] Horowitz, L. M. (1997). The circumplex structure of interpersonal problems. *Circumplex models of personality and emotions*, (January 1997):347–384.

[Huang and Khan, 2017] Huang, Y. and Khan, S. M. (2017). DyadGAN : Generating Facial Expressions in Dyadic Interactions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–18.

[Isbister, 2006] Isbister, K. (2006). *Better Game Characters by Design: A Psychological Approach*. Series in edition.

[Ishii, 2010] Ishii, R. (2010). An Empirical Study of Eye-gaze Behaviors : Towards the Estimation of An Empirical Study of Eye-gaze Behaviors : Towards the Estimation of Conversational Engagement in Human-Agent Communication. (January).

[Jack et al., 2014] Jack, R. E., Garrod, O. G., and Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology*, 24(2):187–192.

[Janssoone, 2015] Janssoone, T. (2015). Temporal Association Rules for Modelling Multimodal Social Signals. pages 575–579.

[Janssoone, 2016] Janssoone, T. (2016). Using temporal association rules for the synthesis of embodied conversational agents with a specific stance. In *16th International Conference of Intelligent Virtual Agents*, pages 175–189, Los Angeles, CA, USA.

[Kam and Fu, 2000] Kam, P.-s. and Fu, A. W.-c. (2000). Discovering Temporal Patterns for Interval-based Events. *Lecture Notes in Computer Science*, 1874(5):317–326.

[Kasap et al., 2009] Kasap, Z., Moussa, M. B., Chaudhuri, P., and Magnenat-Thalmann, N. (2009). Making them remember - Emotional virtual characters with memory. *IEEE Computer Graphics and Applications*, 29(2):20–29.

[Keating and al, 1981] Keating, C. F. and al, E. (1981). Culture and the perception of social dominance from facial expression. *Journal of Personality and Social Psychology*, 40(4):615–626.

[Keltner, 1995] Keltner, D. (1995). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*, 68(3):441–454.

[Kendon, 1967] Kendon, A. (1967). Some functions of gaze-direction in social inter- action. *Acta psychologica*, 26(1):22–63.

[Kiesler, 1996] Kiesler (1996). *Contemporary Interpersonal Theory and Research, Personality, Psychopathol- ogy and Psychotherapy*.

[Knutson, 1996] Knutson, B. (1996). Facial expressions of emotion influence interpersonal trait inferences. *Journal of Nonverbal Behavior*, 20(3):165–182.

[Konijn and Hoorn, 2005] Konijn, E. A. and Hoorn, J. F. (2005). Some like it bad: Testing a model for perceiving and experiencing fictional characters. *Media Psychology*, 7(2):107–144.

[Lafrance, 1982] Lafrance, M. (1982). *Posture Mirroring and Rapport*. Human Sciences Press.

[Leary, 1957] Leary, T. (1957). *Interpersonal Diagnosis of Personality: Functional Theory and Methodology for Personality Evaluation*. Ronald Press, New York.

[Lee and Marsella, 2011] Lee, J. and Marsella, S. (2011). Modeling Side Participants and Bystanders : the Importance of Being a Laugh Track. In *Intelligent Virtual Agents*, volume 6895, pages 240–247.

[Lee and Marsella, 2012] Lee, J. and Marsella, S. (2012). Modeling speaker behavior: A comparison of two approaches. In *The 12th International Conference on Intelligent Virtual Agents (IVA 2012)*, pages 161–174.

[Lee and Marsella, 2010] Lee, J. and Marsella, S. C. (2010). Predicting speaker head nods and the effects of affective information. *IEEE Transactions on Multimedia*, 12(6):552–562.

[Locke, 2000] Locke, K. (2000). Journal of Personality Development and Validation of a Scale to Measure Perceived Control of Internal States. *Journal of personality assessment*, 75:249–267.

[Locke, 2012] Locke, K. D. (2012). Circumplex Measures of Interpersonal Constructs. *Handbook of Interpersonal Psychology: Theory, Research, Assessment, and Therapeutic Interventions*, (March):313–324.

[Locke and Adamic, 2012] Locke, K. D. and Adamic, E. J. (2012). Interpersonal circumplex vector length and interpersonal decision making. *Personality and Individual Differences*, 53(6):764–769.

[Locke and Sadler, 2007] Locke, K. D. and Sadler, P. (2007). Self-efficacy, values, and complementarity in dyadic interactions: Integrating interpersonal and social-cognitive theory. *Personality and Social Psychology Bulletin*, 33(1):94–109.

[Mackenzie et al., 1985] Mackenzie, A. S., Beaumont, C., Boutilier, R., Rullkotter, J., Murrell, S. A. F., Mason, R., Eglinton, G., and McKenzie, D. P. (1985). The Aromatization and Isomerization of Hydrocarbons and the Thermal and Subsidence History of the Nova Scotia Margin. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 315(1531):203–232.

[Magnusson, 2000] Magnusson, M. S. (2000). Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior research methods, instruments, & computers*, 32(1):93–110.

[Mancini et al., 2019] Mancini, M., Biancardi, B., Dermouche, S., Lerner, P., and Pelachaud, C. (2019). Managing Agent ' s Impression Based on User ' s Engagement Detection. In *IVA*.

[Mancini and Pelachaud, 2007] Mancini, M. and Pelachaud, C. (2007). Dynamic Behavior Qualifiers for Conversational Agents. In *Intelligent Virtual Agents*, pages 112–124.

[Mariooryad and Busso, 2013] Mariooryad, S. and Busso, C. (2013). Analysis and Compensation of the Reaction Lag of Evaluators in Continuous Emotional Annotations. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII 2013)*, pages 85–90.

[Mariooryad and Busso, 2015] Mariooryad, S. and Busso, C. (2015). Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. *IEEE Transactions on Affective Computing*, 6(2):97–108.

[Martínez and Yannakakis, 2011] Martínez, H. and Yannakakis, G. (2011). Mining multimodal sequential patterns: a case study on affect detection. In *In Proceedings of the 13th International Conference on Multimodal Interfaces*, pages 3–10.

[Mashat et al., 2013] Mashat, A. F., Fouad, M. M., and Yu, P. S. (2013). Discovery of Association Rules from University Admission System Data. (May):1–7.

[Mcdaniel and Si, 2014] Mcdaniel, J. D. and Si, M. (2014). Length of Smile Apex as Indicator of Faked Expression. In *Affective Agents Workshop at IVA*, pages 25–32.

[McKeown et al., 2012] McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schröder, M. (2012). The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17.

[McNeill, 1992] McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Uni- versity of Chicago press.

[Mehrabian, 1969] Mehrabian, A. (1969). Significance of posture and position in the communication of attitude and status relationships. *Psychological Bulletin*, 71(5):359–372.

[Mihoub et al., 2016] Mihoub, A., Bailly, G., Wolf, C., and Elisei, F. (2016). Graphical models for social behavior modeling in face-to face interaction. *Pattern Recognition Letters*, 74(1):82–89.

[Mörchen, 2007] Mörchen, F. (2007). Unsupervised pattern mining from symbolic temporal data. *ACM SIGKDD Explorations Newsletter*, 9(1):41–55.

[Mota and Picard, 2003] Mota, S. and Picard, R. W. (2003). Automated Posture Analysis for Detecting Learner's Interest Level. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 5:1–6.

[Murphy-chutorian et al., 2009] Murphy-chutorian, E., Member, S., and Trivedi, M. M. (2009). Head Pose Estimation in Computer Vision :. *Analysis*, 31(4):607–626.

[Nakagaito et al., 2009] Nakagaito, F., Ozaki, T., and Ohkawa, T. (2009). Discovery of quantitative sequential patterns from event sequences. In *IEEE International Conference on Data Mining Workshops*, number 1, pages 31–36, Pisa, Italy.

[Nakano, 2015] Nakano, Y. I. (2015). Predicting Participation Styles using Co-occurrence Patterns of Nonverbal Behaviors in Collaborative Learning. pages 91–98.

[Nakano, Yukiko I. and Ishii, 2010] Nakano, Yukiko I. and Ishii, R. (2010). Estimating User's Engagement from Eye-gaze Behaviors in Human-agent Conversations. In *The 5th International Conference on Intelligent User Interfaces*, pages 139—-148. ACM.

[Nguyen et al., 2014] Nguyen, L. S., Frauendorfer, D., Mast, M. S., and Gatica-perez, D. (2014). Hire me : Computational inference of hirability in employment interviews based on nonverbal behavior. (September).

[Niewiadomski et al., 2011] Niewiadomski, R., Hyniewska, S. J., and Pelachaud, C. (2011). Constraint-based model for synthesis of multimodal sequential expressions of emotions. *IEEE Transactions on Affective Computing*, 2(3):134–146.

[Nitesh V. Chawla et al., 2002] Nitesh V. Chawla, Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357.

[Nojavanasghari et al., 2016] Nojavanasghari, B., Baltru, T., Hughes, C. E., and Morency, L.-p. (2016). The Future Belongs to the Curious : Towards Automatic Understanding and Recognition of Curiosity in Children The Future Belongs to the Curious : Towards Automatic Understanding and Recognition of Curiosity in Children. (September).

[Nojavanasghari et al., 2018] Nojavanasghari, B., Huang, Y., and Khan, S. (2018). Interactive Generative Adversarial Networks for Facial Expression Generation in Dyadic Interactions.

[Nojavanasghari and Hughes, 2017] Nojavanasghari, B. and Hughes, C. E. (2017). Exceptionally Social : Design of an Avatar-Mediated Interactive System for Promoting Social Skills in Children with Autism. (May).

[Nomura et al., 2006] Nomura, T., Kanda, T., and Suzuki, T. (2006). Experimental Investigation into Influence of Negative Attitudes toward Robots on Human–Robot Interaction. *AI & Society*, 20(2).

[Ochs et al., 2010] Ochs, M., Sabouret, N., and Corruble, V. (2010). Simulation of the Dynamics of Nonplayer Characters ' Emotions and Social Relations in Games. (January).

[Ojanen et al., 2005] Ojanen, T., Gro, M., and Salmivalli, C. (2005). An Interpersonal Circumplex Model of Children ' s Social Goals : Links With Peer-Reported Behavior and Sociometric Status. 41(5):699–710.

[Op Den Akker et al., 2013] Op Den Akker, R., Bruijnes, M., Peters, R., and Krikke, T. (2013). Interpersonal stance in police interviews: Content analysis. *Computational Linguistics in the Netherlands Journal*, 3:193–216.

[Ortony and Clore, 1988] Ortony, A. and Clore, G. (1988). Review Reviewed Work (s): The Cognitive Structure of Emotions . by Andrew Ortony , Gerald L . Clore and Allan Collins Review by : B . N . Colby Published by : American Sociological Association Stable URL : http://www.jstor.org/stable/2074241. (January).

[Otsuka et al., 2007] Otsuka, K., Sawada, H., and Yamato, J. (2007). Automatic inference of cross-modal nonverbal interactions in multiparty conversations. In *9th international conference on Multimodal interfaces (ICMI'07)*, pages 255–262.

[Pandzic and Forchheimer,] Pandzic, I. S. and Forchheimer, R. *MPEG-4 Facial Animation The Standard , Implementation*. John Wiley & Sons.

[Pecune, 2016] Pecune, F. (2016). Modélisation de la prise de décision d ' un agent conversationnel animé en fonction de son attitude sociale.

[Pecune et al., 2014] Pecune, F., Cafaro, A., Chollet, M., Philippe, P., and Pelachaud, C. (2014). Suggestions for Extending SAIBA with the VIB Platform. In *Workshop on Architectures and Standards for IVAs, held at the '14th International Conference on Intelligent Virtual Agents (IVA 2014)*, pages 16–20, Boston, MA, USA. Bielefeld eCollections.

[Pecune et al., 2016] Pecune, F., Ochs, M., Marsella, S., and Pelachaud, C. (2016). SOCRATES: from SOCial Relation to ATtitude ExpressionS. *Conference on Autonomous Agents & Multiagent Systems (AAMAS)*, pages 921–930.

[Pei et al., 2001] Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., and Hsu, M.-C. (2001). PrefixSpan,: mining sequential patterns efficiently by prefix-projected pattern growth. In *17th International Conference on Data Engineering (ICDE 2001)*, pages 215–224.

[Peters et al., 2005] Peters, C., Pelachaud, C., Bevacqua, E., Mancini, M., and Poggi, I. (2005). Engagement Capabilities for ECAs. *AAMAS'05 workshop Creating Bonds with ECAs*.

[Pickering and Garrod, 2004] Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *The Behavioral and brain sciences*, 27(2):169–190; discussion 190–226.

[Poggi, 2007] Poggi, I. (2007). *Mind, Hands, Face and Body: A Goal and Belief View of Multimodal Communication*.

[Rajendran, 2004] Rajendran, G. (2004). Mind reading: An interactive guide to emotions. *Autism*, 8(3):341–343.

[Rakesh Agrawal, 1994] Rakesh Agrawal, R. S. (1994). Fast Algorithms for Mining Association Rules. *The Annals of pharmacotherapy*, 42(1):62–70.

[Ravenet et al., 2015] Ravenet, B., Cafaro, A., Biancardi, B., Ochs, M., and Pelachaud, C. (2015). Conversational behavior reflecting interpersonal attitudes in small group interactions. In *15th International Conference of Intelligent Virtual Agents (IVA 2015)*, pages 375–388.

[Ravenet et al., 2013] Ravenet, B., Ochs, M., and Pelachaud, C. (2013). From a user-created corpus of virtual agent's non-verbal behavior to a computational model of interpersonal attitudes. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8108 LNAI:263–274.

[Richmond, V. & McCroskey, 2000] Richmond, V. & McCroskey, J. (2000). *Nonverbal Behavior in Interpersonal Relations*. Allyn and Bacon.

[Rogelberg, 2017] Rogelberg, S. G. (2017). Self-Efficacy. In *The SAGE Encyclopedia of Industrial and Organizational Psychology, 2nd edition*.

[Rosenthal-von der Pütten et al., 2019] Rosenthal-von der Pütten, A. M., Straßmann, C., Yaghoubzadeh, R., Kopp, S., and Krämer, N. C. (2019). Dominant and submissive nonverbal behavior of virtual agents and its effects on evaluation and negotiation outcome in different age groups. *Computers in Human Behavior*, 90(August 2018):397–409.

[Ruan et al., 2014] Ruan, G., Zhang, H., and Plale, B. (2014). Parallel and Quantitative Sequential Pattern Mining for Large-scale Interval-based Temporal Data. In *2014 IEEE International Conference on Big Data (BigData 2014)*, pages 32–39.

[Ruiz et al., 2010] Ruiz, M., Pincus, A. L., Borkovec, T. D., Echemendia, R., Castonguay, L. G., and Ragusea, S. (2010). Validity of the IIP for Predicting Treatment Outcome: An Investigation with the Pennsylvania Practice Research Network. *Journal of Personality Assessment*, 83(3):332–344.

[Sanghvi et al., 2011] Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., and Paiva, A. (2011). Automatic analysis of affective postures and body motion to detect engagement with a game companion. *Proceedings of the 6th international conference on Human-robot interaction - HRI '11*, page 305.

[Sariyanidi et al., 2015] Sariyanidi, E., Gunes, H., and Cavallaro, A. (2015). Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1113–1133.

[Scherer, 2005] Scherer, K. R. (2005). What are emotion? And how can they be measured? *Social Science Information*, 44(4):695–729.

[Schmidt, J. A., Wagner, C. C., & Kiesler, 1999] Schmidt, J. A., Wagner, C. C., & Kiesler, D. J. (1999). Psychometric and circumplex properties of the octant scale Impact Message Inventory. *Journal of Counseling Psychology*, 46(3):325–334.

[Schutz, 1958] Schutz, W. C. (1958). Firo. A Three-Dimensional Theory of Interpersonal Behavior. *International Journal of Group Psychotherapy*, 10(3):360.

[Sidner et al., 2005] Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., and Rich, C. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164.

[Sidner et al., 2003] Sidner, C. L., Lee, C., and Lesh, N. (2003). Engagement by Looking: Behaviours for Robots when Collaborating with People. *Proceedings of DiaBruck (the 7th Workshop on Semantics and Pragmatics of Dialogue)*, pages 123–130.

[Sodano and Tracey, 2007] Sodano, S. M. and Tracey, T. J. G. (2007). Interpersonal Traits in Childhood : Development of the Child and Adolescent Interpersonal Traits in Childhood : Development of the Child and Adolescent Interpersonal Survey. (January).

[Srikant and Agrawal, 1996] Srikant, R. and Agrawal, R. (1996). Mining Sequential Patterns: Generalizations and Performance Improvements. In *5th International Conference on Extending Database Technology: Advances in Database Technology (EDBT '96)*, pages 3–17.

[Srivastava et al., 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfittin. *Journal of Machine Learning Research*, 15:1929–1958.

[Stivers, 2008] Stivers, T. (2008). Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation. *Research on Language and Social Interaction*, 41(1):31–57.

[Straßmann et al., 2016] Straßmann, C., von der Pütten, A. R., and Yaghoubzadeh, R. (2016). The effect of an intelligent virtual agent' s nonverbal behavior with regard to dominance and cooperativity Dominant nonverbal behavior. *Proceedings of International Conference on Intelligent Virtual Agents*.

[Sullivan, 1953] Sullivan, H. S. (1953). *The interpersonal theory of psychiatry.* New york: edition.

[Szekrényes, 2014] Szekrényes, I. (2014). Annotation and interpretation of prosodic data in the HuComTech corpus formultimodal user interfaces. *Journal on Multimodal User Interfaces*, 2(8):143–150.

[Tian et al., 2017] Tian, X., Zhang, J., Ma, Z., He, Y., Wei, J., Wu, P., Situ, W., Li, S., and Zhang, Y. (2017). Deep lstm for large vocabulary continuous speech recognition. *ArXiv*.

[Tiedens et al., 2000] Tiedens, L., Ellsworth, P., and Mesquita, B. (2000). Stereotypes about sentiments and status: Emotional expectations for high- and low-status group members. *Personality and Social Psychology Bulletin*, 26(5):560–574.

[Trobst, 2000] Trobst, K. K. (2000). An interpersonal conceptualization and quantification of social support transactions. *Personality and Social Psychology*, (Bulletin 26):971–986.

[Vahdatpour et al., 2005] Vahdatpour, A., Amini, N., and Sarrafzadeh, M. (2005). Toward Unsupervised Activity Discovery Using Multi-Dimensional Motif Detection in Time Series. pages 1261–1266.

[van Vugt et al., 2006] van Vugt, H. C., Hoorn, J. F., Konijn, E. A., and de Bie Dimitriadou, A. (2006). Affective affordances: Improving interface character engagement through interaction. *International Journal of Human Computer Studies*, 64(9):874–888.

[van Waterschoot et al., 2018] van Waterschoot, J., Bruijnes, M., Flokstra, J., Reidsma, D., Davison, D., Theune, M., and Heylen, D. (2018). Flipper 2.0. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, number November, pages 43–50.

[Vilhjalmsson et al., 2007] Vilhjalmsson, H., Cantelmo, N., Cassell, J., Chafai, N. E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A. N., Pelachaud, C., Ruttkay, Z., Thórisson, K. R., Van Welbergen, H., and Van Der Werf, R. J. (2007). The behavior markup language: Recent developments and challenges. *Lecture Notes in Artificial Intelligence*, 4722(1):99–111.

[Vinciarelli et al., 2009] Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing : Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759.

[Volpe et al., 2016] Volpe, G., Alborno, P., Camurri, A., Coletta, P., Ghisio, S., Mancini, M., Niewiadomski, R., and Piana, S. (2016). Designing Multimodal Interactive Systems Using EyesWeb XMI. *SERVE@AVI*, pages 49–56.

[Wang et al., 2018] Wang, N., Gao, X., Tao, D., Yang, H., and Li, X. (2018). Facial feature point detection: A comprehensive survey. *Neurocomputing*, 275:50–65.

[Wiggins, 1979] Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology*, (37):395–412.

[Wiggins, 1995] Wiggins, J. S. (1995). *IAS, Interpersonal Adjective Scales Professional Manual*. Psychologi edition.

[Wiggins et al., 1988] Wiggins, J. S., Trapnell, P., and Phillips, N. (1988). Psychometric and Geometric Characteristics of the Revised Interpersonal Adjective Scales (IAS-R).

[With and Kaiser, 2011] With, S. and Kaiser, S. (2011). Sequential patterning of facial actions in the production and perception of emotional expressions. *Swiss Journal of Psychology*, 70(4):241–252.

[Wittenburg et al., 2006] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. *Proceedings Language Resources and Evaluation*, pages 1556–1559.

[Woolf and Burleson, 2009] Woolf, B. and Burleson, W. (2009). Affect-aware tutors : recognising and responding to student affect Ivon Arroyo , Toby Dragon and David Cooper Rosalind Picard. 4.

[Yannakakis, 2018] Yannakakis, G. N. (2018). The Ordinal Nature of Psychophysiology. In *Proceedings of the 5th International Conference on Physiological Computing Systems*, Seville, Spain.

[Yannakakis et al., 2017] Yannakakis, G. N., Cowie, R., and Busso, C. (2017). The Ordinal Nature of Emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 248–255.

[Yannakakis and Martinez, 2015] Yannakakis, G. N. and Martinez, H. P. (2015). Grounding truth via ordinal annotation. *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, pages 574–580.

[Yu et al., 2004a] Yu, C., Aoki, P. M., and Woodruff, A. (2004a). Detecting User Engagement in Everyday Conversations. *Science*, page 4.

[Yu et al., 2004b] Yu, C., Aoki, P. M., and Woodruff, A. (2004b). Detecting User Engagement in Everyday Conversations. (May).

[Yu et al., 2010] Yu, C., Scheutz, M., and Schermerhorn, P. (2010). Investigating multimodal real-time patterns of joint attention in an HRI word learning task. In *5th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2010)*, pages 309–316.

[Yu et al., 2016] Yu, Z., He, X., Black, A. W., and Rudnicky, A. I. (2016). User Engagement Study with Virtual Agents Under Different Cultural Contexts. In *Intelligent Virtual Agents - 16th International Conference, IVA2016*, pages 364–368, Los Angeles, CA, USA.

[Yzerbyt et al., 2008] Yzerbyt, V. Y., Kervyn, N., and Judd, C. M. (2008). Compensation versus halo: The unique relations between the fundamental dimensions of social judgment. *Personality and Social Psychology Bulletin*, 34(8):1110–1123.

[Zhang and Boyles, 2013] Zhang, H. and Boyles, M. J. (2013). Visual exploration and analysis of human-robot interaction rules. In *Visualization and Data Analysis (SPIE 8654)*.

[Zhang et al., 2010] Zhang, H., Fricker, D., Smith, T. G., and Yu, C. (2010). Real-time adaptive behaviors in multimodal human-avatar interactions. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010)*, pages 1–8.

[Zhao et al., 2016] Zhao, R., Sinha, T., Black, A. W., and Cassell, J. (2016). Socially-Aware Virtual Agents : Automatically Assessing Dyadic Rapport from Temporal Patterns of Behavior. In *16th International Conference of Intelligent Virtual Agents*, pages 218–233.

Appendices

Appendix A

Results of the First Study

(1) “strongly disagree”, (2) “partially disagree”, (3) “neutral”, (4) “partially agree”, (5) “strongly agree”.

Variable	M	SD	(1)	(2)	(3)	(4)	(5)
Aggressive	1.93	1.12	**46%**	28%	18%	3%	3%
arrogant	1.93	0.85	34%	**43%**	15%	6%	0%
Assertive	2.93	0.92	6%	15%	**46%**	31%	0%
Compete	2.18	1.04	25%	**31%**	**31%**	12%	0%
Cooperative	3	0.89	3%	15%	34%	**43%**	3%
Depend	2.43	0.96	16%	25%	**50%**	9%	0%
Defiant	2.18	0.91	**34%**	31%	31%	3%	0%
Distant	3.25	0.85	6%	28%	**37%**	28%	0%
Forceful	2.12	0.95	25%	18%	18%	**37%**	0%
Helpful	2.93	1.18	9%	21%	18%	**43%**	6%
Cheerful	2.31	0.79	9%	**34%**	31%	25%	0%
Timid	2.25	1.18	28%	25%	**34%**	9%	3%
Unauthorative	2.68	1.07	9%	31%	**34%**	18%	6%
tender	2.56	0.96	6%	12%	**43** %	31%	6%
Withdrawn	2.56	1.09	12%	18%	**46%**	21%	0%
Leader-like	2.50	0.96	9%	**34%**	28%	28%	0%

Table A.1 – Mean, standard deviation of variables, and distribution of participants’ answers for *DomRef*.

Variable	M	SD	(1)	(2)	(3)	(4)	(5)
Aggressive	1.56	0.91	**62%**	25%	9%	0%	2%
arrogant	2.15	0.98	**31%**	**31%**	28%	9%	0%
Assertive	3.03	0.99	6%	21%	**40%**	25%	6%
Compete	2.43	1.07	15%	**46%**	21%	9%	6%
Cooperative	3.34	0.82	3%	6%	**50%**	34%	6%
Depend	2.59	0.91	15%	21%	**50%**	12%	0%
Defiant	2.76	0.88	25%	25%	**37%**	6%	6%
Distant	2.81	1.06	12%	21%	**43%**	15%	6%
Forceful	2.28	0.99	21%	*40*%	28%	6%	3%
Helpful	3.37	1.15	9%	12%	21%	**43%**	12%
Cheerful	2.62	1%	12%	**34%**	**34%**	15%	3%
Timid	2.73	0.73	18%	31%	**43%**	6%	0%
Unauthorative	2.59	0.97	15%	**37%**	28%	18%	0%
tender	2.68	1.09	18%	18%	**40%**	18%	3%
Withdrawn	2.65	0.78	9%	25%	**56%**	9%	0%
Leader-like	2.56	1.29	**28%**	21%	21%	21%	6%

Table A.2 – Mean, standard deviation of variables, and distribution of participants' answers for *FrRef*.

Variable	M	SD	(1)	(2)	(3)	(4)	(5)
Aggressive	2.20	0.74	12%	15%	21%	**39%**	10%
arrogant	3.42	0.79	9%	12%	18%	**45%**	14%
Assertive	3	0.79	4%	28%	**35%**	25%	6%
Compete	3.32	0.77	7%	20%	21%	**31%**	18%
Cooperative	2.29	0.78	21%	**37%**	31%	7%	1%
Depend	2.68	0.69	9%	26%	**51%**	10%	1%
Defiant	3.26	0.81	9%	10%	**35%**	31%	12%
Distant	3.46	0.8	7%	7%	**32%**	**32%**	18%
Forceful	3.32	0.83	7%	17%	26%	**31%**	17%
Helpful	2.52	0.73	10%	**39%**	32%	17%	0%
Cheerful	2.18	0.65	26%	**39%**	23%	10%	0%
Timid	2.64	0.77	21%	25%	**29%**	14%	9%
Unauthorative	2.76	0.81	15%	**29%**	25%	21%	7%
tender	2.12	0.8	29%	**40%**	20%	6%	3%
Withdrawn	2.65	0.74	12%	25%	**46%**	15%	0%
Leader-like	3.1	0.66	12%	12%	32%	**35%**	6%

Table A.3 – Mean, standard deviation and frequency of participants' answers for *DomInc*.

Variable	M	SD	(1)	(2)	(3)	(4)	(5)
Aggressive	2.14	0.74	**40%**	17%	31%	9%	1%
arrogant	2.29	0.72	**35%**	20%	25%	15%	3%
Assertive	3.03	0.50	9%	20%	**34%**	29%	6%
Compete	2.48	0.82	**29%**	20%	25%	21%	3%
Cooperative	3.03	0.59	3%	28%	**39%**	21%	7%
Depend	2.76	0.60	10%	21%	**48%**	17%	1%
Defiant	2.92	0.87	**32%**	28%	20%	14%	4%
Distant	3.06	0.73	10%	18%	29%	**34%**	6%
Forceful	2.28	0.71	28%	25%	**37%**	9%	0%
Helpful	2.90	0.67	1%	**37%**	34%	21%	4%
Cheerful	2.82	0.80	10%	**31%**	25%	29	3%
Timid	2.95	0.71	17%	14%	**29%**	34%	4%
Unauthorative	3.18	0.58	6%	18%	31%	**37%**	6%
tender	2.76	0.71	7%	32%	**39%**	15%	4%
Withdrawn	2.81	0.62	10%	29%	**31%**	23%	4%
Leader-like	2.62	0.62	14%	**35%**	26%	20%	3%

Table A.4 – Mean, standard deviation and frequency of participants' answers for *DomDec* in the first experiment.

Variable	M	SD	(1)	(2)	(3)	(4)	(5)
Aggressive	2.01	0.59	**34%**	**34%**	26%	4%	0%
arrogant	2.21	0.55	25%	**35%**	31%	7%	0%
Assertive	3.04	0.57	3%	21%	43%	**29%**	1%
Compete	2.56	0.64	10%	**42%**	28%	17%	1%
Cooperative	3.34	0.67	1%	17%	34%	**39%**	7%
Depend	2.76	0.59	12%	14%	**67%**	6%	0%
Defiant	2.31	0.71	26%	25%	**40%**	6	1%
Distant	2.82	0.48	9%	28%	**39%**	17%	6%
Forceful	2.60	0.57	12%	28%	**46%**	10%	1%
Helpful	3.26	0.76	0%	26%	**31%**	**31%**	10%
Cheerful	2.78	0.49	4%	35%	**37%**	20%	1%
Timid	2.51	0.71	20%	26%	**35%**	15%	1%
Unauthoritative	2.73	0.39	4%	29%	**54%**	9%	1%
tender	3.1	0.49	1%	15%	**53%**	29%	0%
Withdrawn	2.82	0.48	4%	21%	**59%**	14%	0%
Leader-like	2.7	0.49	9%	29%	**42%**	18%	0%

Table A.5 – Mean, standard deviation and frequency of participants' answers of variables for *FrInc*.

Variable	M	SD	(1)	(2)	(3)	(4)	(5)
Aggressive	2.95	0.65	17%	12%	**34%**	29%	6%
arrogant	2.92	0.53	10%	20%	**39%**	25%	4%
Assertive	2.98	0.43	4%	26%	**35%**	31%	1%
Compete	2.92	0.48	4%	29%	**37%**	25%	3%
Cooperative 2.85	0.65	10%	20%	**45%**	18%	4%	
Depend	2.79	0.67	12%	15%	**53%**	17%	1%
Defiant	3.17	0.60	9%	14%	**34%**	**34%**	7%
Distant	3.18	0.68	9%	14%	29%	**42%**	4%
Forceful	3.18	0.47	3%	18%	35%	**40%**	1%
Helpful	2.87	0.80	10%	23%	**34%**	29%	1%
Cheerful	2.53	0.64	20%	21%	**42%**	15%	0%
Timid	2.35	0.82	29%	23%	**31%**	12%	3%
Unauthoritative	2.53	0.81	21%	21%	**40%**	12%	3%
Tender	2.54	0.67	20%	20%	**43%**	15%	0%
Withdrawn	2.43	0.60	21%	17%	**54%**	6%	0%
Leader-like	3.01	0.52	7%	15%	**45%**	29%	1%

Table A.6 – Mean, standard deviation and frequency of participants' answers of variables for *FrDec* in the first experiment.

Appendix B

Results of the Second Study

Variable	M	SD	(1)	(2)	(3)	(4)	(5)
Aggressive	2.48	0.639	22%	27%	**31%**	15%	3%
arrogant	2.54	0.72	18%	**31%**	**31%**	14%	4%
Assertive	3.31	0.35	5%	0%	30%	**61%**	3%
Compete	3.16	0.72	7%	22%	28%	**30%**	11%
Cooperative	3.30	0.47	0%	16%	**43%**	32%	7%
Depend	2.85	0.54	3%	21%	**65%**	7%	3%
Defiant	2.68	0.75	11%	26%	**48%**	9%	4%
Distant	2.86	0.51	4%	33%	**37%**	20%	4%
Forceful	3	0.59	3%	29%	**37%**	23%	6%
Helpful	3.40	0.30	0%	6%	**51%**	38%	4 %
Cheerful	3.03	0.47	2%	26%	**42%**	23%	5 %
Timid	2.08	1.02	**38%**	33%	13%	9%	5%
Unauthorit.	2.7	0.49	2%	38%	**47%**	9%	2%
Tender	2.91	0.54	5%	**31%**	**31%**	29%	2%
Withdrawn	2.82	0.64	7%	16%	**65%**	9%	2%
Leader-like	3.13	0.54	2%	5%	30%	**46%**	15%
Dominant	3.68	0.60	4%	25%	**33%**	26%	10%
Submissive	2.71	0.60	4%	36%	**45%**	10%	3%
Hostile	2.47	0.75	14%	**39%**	33%	8%	4%
Friendly	3.31	0.37	1%	24%	28%	**33%**	12%

Table B.1 – Mean, standard deviation and frequency of participants' answers for the seven reference videos.

Variable	M	SD	(1)	(2)	(3)	(4)	(5)
Aggressive	3.69	0.40	1%	7%	26%	**52%**	13%
arrogant	3.66	0.41	0%	7%	33%	**44%**	14%
Assertive	3.52	0.67	0%	16%	29%	**39%**	14%
Compete	3.74	0.50	0%	6%	27%	**51%**	14%
Cooperative	2.54	0.36	5%	**45%**	39%	8%	1%
Depend	2.9	0.37	0%	20%	**69%**	9%	1%
Defiant	3.63	0.34	0%	8%	29%	**53%**	9%
Distant	3.47	0.54	0%	11%	**41%**	34%	12%
Forceful	3.72	0.44	0%	10%	22%	**52%**	15%
Helpful	2.67	0.46	5%	42%	**34%**	14%	3%
Cheerful	2.47	0.46	10%	**44%**	32%	11%	1%
Timid	2.46	0.45	7%	**45%**	40%	5%	1%
Unauthoritative	2.76	0.72	8%	32%	**35%**	21%	2%
Tender	2.38	0.47	14%	**48%**	23%	10%	3%
Withdrawn	3.19	0.52	1%	16%	**52%**	23%	7%
Leader-like	3.43	0.58	2%	16%	26%	**45%**	9%
Dominant	3.79	0.42	0%	6%	25%	**51%**	17%
Submissive	2.53	0.60	8%	**42%**	36%	12%	0%
Hostile	3.6	0.42	2%	7%	32%	**44%**	13%
Friendly	2.43	0.45	12%	**45%**	28%	12%	1%

Table B.2 – Mean, standard deviation and frequency of participants' answers of variables for *DomInc*.

Variable	M	SD	(1)	(2)	(3)	(4)	(5)
Aggressive	2.90	0.89	13%	17%	**42%**	18%	8%
arrogant	2.88	0.89	13%	16%	**46%**	15%	8%
Assertive	2.96	0.95	15%	9%	**44%**	24%	6%
Compete	3.16	1.03	14%	3%	**40%**	35%	6%
Cooperative	2.95	0.85	15%	12%	**40%**	24%	7%
Depend	2.72	0.92	13%	23%	**48%**	6%	8%
Defiant	2.93	0.8	14%	14%	**41%**	22%	7%
Distant	2.9	0.92	14%	23%	**27%**	26%	8%
Forceful	2.96	0.90	13%	14%	**42%**	21%	8%
Helpful	3.07	0.74	14%	11%	**35%**	30%	8%
Cheerful	2.94	0.90	17%	**31%**	15%	28%	7%
Timid	2.62	0.97	15%	33%	**35%**	4%	11%
Unauthoritative	2.75	0.85	15%	21%	**40%**	17%	5%
Tender	2.85	0.82	14%	25%	**27%**	25%	7%
Withdrawn	2.88	0.88	13%	8%	**63%**	7%	8%
Leader-like	3.11	1.01	17%	4%	**35%**	**35%**	7%
Dominant	3.06	0.84	14%	10%	**40%**	24%	10%
Submissive	2.90	0.88	13%	10%	**57%**	11%	8%
Hostile	2.8	0.88	13%	25%	**36%**	16%	8%
Friendly	2.97	0.93	13%	22%	**27%**	26%	10%

Table B.3 – Mean, standard deviation and frequency of participants' answers of variables for *DomDec*

Variable	M	SD	(1)	(2)	(3)	(4)	(5)
Aggressive	2.97	0.29	1%	21%	**57%**	20%	0%
arrogant	2.95	0.23	0%	20%	**64%**	14%	1%
Assertive	3.5	0.34	0%	7%	37%	**52%**	3%
Compete	3.42	0.30	1%	5%	44%	**47%**	1%
Cooperative	3.21	0.34	1%	5%	**61%**	32%	0%
Depend	2.7	0.21	0%	31%	**68%**	0%	0%
Defiant	2.9	0.15	0%	18%	**70%**	11%	0%
Distant	2.82	0.25	1%	28%	**58%**	11%	1%
Forceful	3.04	0.35	1%	20%	**52%**	26%	0%
Helpful	3.36	0.33	0%	4%	**52%**	42%	1%
Cheerful	3.09	0.22	0%	14%	**62%**	22%	1%
Timid	2.54	0.29	0%	**52%**	44%	3%	0%
Unauthoritative	2.63	0.41	2%	**46%**	41%	9%	0%
Tender	3.16	0.23	0%	17%	**47%**	33%	1%
Withdrawn	2.95	0.15	0%	6%	**91%**	2%	0%
Leader-like	3.60	0.30	0%	0%	35%	**63%**	1%
Dominant	3.35	0.32	0%	7%	**48%**	42%	1%
Submissive	2.75	0.21	0%	30%	**67%**	2%	0%
Hostile	2.94	0.26	1%	28%	**47%**	21%	1%
Friendly	3.17	0.26	1%	17%	**40%**	**40%**	0%

Table B.4 – Mean, standard deviation and frequency of participants' answers of variables for *FrInc*.

Variable	M	SD	(1)	(2)	(3)	(4)	(5)
Aggressive	3.19	0.45	0%	17%	**47%**	32%	2%
arrogant	3.14	0.39	0%	15%	**55%**	29%	0%
Assertive	3.19	0.42	0%	13%	**56%**	28%	2%
Compete	3.12	0.54	2%	15%	**51%**	31%	0%
Cooperative	2.98	0.16	0%	11%	**78%**	10%	0%
Depend	2.86	0.23	0%	18%	**76%**	5%	0%
Defiant	3.11	0.32	1%	13%	**59%**	26%	0%
Distant	3.01	0.34	0%	18%	**62%**	19%	0%
Forceful	3.30	0.33	0%	11%	**47%**	39%	1%
Helpful	2.95	0.32	0%	19%	**65%**	15%	0%
Cheerful	3.05	0.32	1%	18%	**55%**	25%	0%
Timid	2.64	0.45	1%	40%	**51%**	7%	0%
Unauthoritative	2.67	0.37	0%	40%	**51%**	8%	0%
Tender	2.86	0.32	0%	29%	**54%**	16%	0%
Withdrawn	2.90	0.19	1%	15%	**75%**	8%	0%
Leader-like	3.54	0.37	0%	10%	27%	**60%**	2%
Dominant	3.38	0.44	1%	6%	**47%**	42%	2%
Submissive	3.72	0.32	0%	33%	**61%**	4%	1%
Hostile	3.03	0.32	0%	19%	**58%**	22%	0%
Friendly	2.80	0.39	0%	35%	**47%**	16%	0%

Table B.5 – Mean, standard deviation and frequency of participants' answers of variables for *FrDec*

Appendix C

Engagement Prediction

		Mean	Level1	Level2	Level3	Level4	Level5
Gaze	Rappel	71%	0%	23%	03%	98%	0%
	Precision	62%	0%	47%	53%	72%	0%
	F-measure	62%	0%	31%	07%	83%	0%
Head rotation	Rappel	72%	0%	28%	6%	98%	0%
	Precision	61%	0%	45%	46%	73%	0%
	F-measure	63%	0%	35%	10%	84%	0%
Au intensities	Rappel	77%	32%	44%	45%	93%	29%
	Precision	76%	61%	54%	54%	83%	68%
	F-measure	76%	42%	48%	49%	88%	41%
Au activation	Rappel	76%	33%	39%	44%	93%	13%
	Precision	74%	69%	55%	53%	81%	61%
	F-measure	73%	45%	46%	48%	86%	22%
Aus	Rappel	92%	78%	90%	82%	96%	69%
	Precision	91%	95%	82%	86%	94%	83%
	F-measure	91%	86%	86%	84%	95%	75%
Smile (AU12)	Rappel	28%	30%	0%	25%	41%	74%
	Precision	45%	75%	0%	23%	73%	10%
	F-measure	30%	43%	0%	24%	53%	17%
All features	Rappel	97%	94%	97%	94%	98%	92%
	Precision	97%	99%	94%	94%	94%	98%
	F-measure	97%	96%	95%	94%	98%	93%

Table C.1 – Prediction of engagement level using several features and expert LSTM.

		Mean	Level1	Level2	Level3	Level4	Level5
Gaze	Rappel	71%	0%	6%	5%	98%	0%
	Precision	60%	0%	34%	50%	71%	0%
	F-measure	60%	0%	11%	09%	83%	0%
Head rotation	Rappel	71%	0	11%	4%	99%	0%
	Precision	61%	0%	4%	54%	72%	0%
	F-measure	61%	0%	17%	8%	83%	0%
Au intensities	Rappel	79%	4%	55%	48%	94%	32%
	Precision	78%	87%	60%	60%	84%	72%
	F-measure	77%	7%	57%	53%	89%	45%
Au activation	Rappel	74%	0%	52%	34%	91%	18%
	Precision	72%	1%	47%	44%	81%	65%
	F-measure	71%	1%	49%	38%	86%	29%
Aus	Rappel	92%	69%	85%	83%	96%	73%
	Precision	92%	93%	85%	86%	94%	86%
	F-measure	92%	79%	85%	84%	95%	79%
All features	Rappel	96%	90%	93%	94%	98%	87%
	Precision	96%	96%	95%	94%	97%	95%
	F-measure	96%	93%	94%	94%	98%	91%

Table C.2 – Prediction of engagement using several novice LSTM.

		Mean	Level1	Level2	Level3	Level4	Level5
Gaze	Rappel	75%	3%	43%	29%	94%	28%
	Precision	73%	66%	50%	54%	80%	66%
	F-measure	72%	6%	46%	38%	86%	40%
Head rotation	Rappel	76%	1%	44%	38%	94%	8%
	Precision	73%	4%	5%	53%	80%	75%
	F-measure	72%	2%	47%	44%	87%	15%
Au intensities	Rappel	94%	84%	79%	89%	97%	83%
	Precision	94%	93%	93%	85%	96%	92%
	F-measure	94%	88%	85%	87%	97%	88%
Au activation	Rappel	92%	82%	84%	92%	94%	80%
	Precision	93%	90%	91%	78%	97%	80%
	F-measure	92%	86%	87%	85%	95%	80%
Aus	Rappel	98%	97%	95%	96%	98%	96%
	Precision	98%	94%	95%	94%	99%	96%
	F-measure	98	96	95	95	98	96
All features	Rappel	99%	93%	97%	95%	99%	97%
	Precision	99%	97%	95%	97%	99%	96%
	F-measure	99%	95%	96%	96%	99%	97%

Table C.3 – Prediction of engagement using dyadic LSTM.