GRAB-HAI : Generating Reciprocally Adaptive Behavior for Human-Agent Interaction

Jieyeon Woo

GRAB-HAI: Generating Reciprocally Adaptive Behavior for Human-Agent Interaction

École Doctorale Sciences Mécaniques, Acoustique, Électronique et Robotique de Paris

THÈSE DE DOCTORAT
DE SORBONNE UNIVERSITÉ

présentée et soutenue publiquement par

Jieyeon Woo

le 11 Décembre 2023

Directrice de thèse: **Catherine Achard**
Co-directrice: **Catherine Pelachaud**

devant le jury composé de :

Mme Ginevra CASTELLANO, Professeure, Uppsala University,	Rapporteure
Mme Magalie OCHS, Maîtresse de conférences, Aix Marseille Université-LIS,	Rapporteure
M. Mohamed CHETOUANI, Professeur, Sorbonne Université-ISIR,	Président
M. Florian PECUNE, Maître de conférences, Université de Bordeaux,	Examinateur
M. Quoc-Cuong PHAM, Directeur de recherche, CEA-LIST,	Examinateur
Mme Catherine ACHARD, Professeure, Sorbonne Université-ISIR,	Directrice de thèse
Mme Catherine PELACHAUD, Directrice de recherche, CNRS-ISIR,	Co-directrice de thèse

Résumé Long

L'information est transférée d'une personne à une autre par le biais de la communication. Par ce transfert, nous transmettons nos pensées et nos intentions par le biais de signaux *multimodaux* tels que les mots, les gestes et la prosodie. Cet échange de signaux est un processus bidirectionnel d'envoi et de réception dans lequel les comportements des interlocuteurs s'adaptent l'un à l'autre. Cette adaptation est continue, dynamique et réciproque, ce que nous appelons l'*adaptation réciproque*. S'adapter aux autres permet aux interactions d'être engageantes et efficaces.

Les agents socialement interactifs (*SIA*; Lugrin et al. [2021]) sont des agents physiques ou virtuels, tels que des robots ou des agents virtuels, capables de mener des conversations naturelles avec des personnes de manière autonome (c.-à-d. leurs utilisateurs) en échangeant des signaux *multimodaux*, verbaux et non verbaux (Ball et al. [2000]) d'une manière socialement intelligente. Comme les interlocuteurs humains, ils peuvent agir en tant que partenaires conversationnels en adaptant leurs comportements en fonction de ceux de leurs interlocuteurs. Le domaine des *SIA*s s'est développé depuis plus de 20 ans sous de nombreux noms tels que agents virtuels intelligents (IVA), agents conversationnels animés (ACAs), et robotique sociale (Cassell [2001], Dautenhahn [1998]). Ils partagent tous une définition similaire et poursuivent le même objectif: faire progresser la recherche et le développement d'agents socialement intelligents montrant des comportements autonomes par le biais d'incarnations physiques ou virtuelles. Les *SIA*s transmettent leur message (intention ou sentiment) verbalement par le biais de mots et émettent un comportement non verbal semblable à celui des humains, tel que des gestes du visage, du corps et des mains pendant le discours (Cassell et al. [2000]), pour compléter leur message.

Les agents animés ont pour objectif commun d'améliorer la communication et l'expérience de l'utilisateur (engagement, rapport et appréciation). Plusieurs travaux se sont concentrés sur l'amélioration de l'interaction elle-même pour une expérience plus fluide avec l'agent (Bailenson and Yee [2005], Huang et al. [2010], Ritschel et al. [2017], Weber et al. [2018]). L'utilisation des *SIA*s a été observée dans diverses applications, montrant leurs utilité dans l'enseignement ou le coaching (Anderson et al. [2013], Pecune et al. [2016], Jones and Castellano [2018], Pereira Santos et al. [2023]), l'assistance (Biancardi et al. [2021], Sidner et al. [2018]) et la santé (Raffard et al. [2018], Ring et al. [2016], Shidara et al. [2022], Khamis et al. [2021]). Dans le domaine des traitements médicaux en particulier, l'utilisation des *SIA*s a considérablement augmenté ces dernières années (Khamis et al. [2021], Shidara et al. [2022].

La recherche sur les *SIA*s s'intéresse à la création d'un *SIA* capable d'interagir avec les gens d'une manière *sociale* et *engageante*. Il est essentiel de garantir les aspects sociaux, d'engagement (le démarrage, le maintien et la fin des liens perçus entre les personnes au cours d'une interaction; Sidner et al. [2003]) et le rapport (affect positif, attention mutuelle et coordination; Tickle-Degnen and Rosenthal [1990]) pour attirer l'attention des utilisateurs et mener l'interaction de manière continue. Pour que la communication soit efficace, les agents jouant le rôle de partenaires d'interaction doivent jouer les rôles actifs de *locuteur* et d'*auditeur*.

Les études sur l'interaction humain-humain ont montré que la communication est *multimodale*. Différents canaux (ou modalités composées de comportements verbaux et non verbaux) de communication participent à la transmission de l'intention et des émotions d'une personne. Ils sont synchronisés à la fois au niveau *intrapersonnel* (entre les indices d'une même personne) et au niveau *interpersonnel* (entre les interlocuteurs). Pour que les *SIA*s jouent le rôle d'interlocuteurs, ils doivent non seulement transmettre leur message en alignant leurs signaux *multimodaux* dérivés d'eux-mêmes (compte tenu de la relation *intrapersonnelle*; Knapp et al. [2013]), mais aussi s'adapter constamment et réciproquement à leurs interlocuteurs, en coordonnant leurs comportements avec les signaux multimodaux de leurs interlocuteurs (compte tenu de la relation *interpersonnelle*; Burgoon et al. [1995]). Démontrer une telle capacité d'adaptation est essentiel pour les relations interpersonnelles (Cappella [1991]) et peut permettre aux *SIA*s d'être perçus comme socials, engageants et établissant un rapport (Biancardi et al. [2021], Oertel et al. [2020], Delaherche and Chetouani [2010], Gupta et al. [2019], Huang et al. [2010], Raffard et al. [2018]).

Objectif de la recherche

La motivation du travail présenté dans cette thèse est de développer un *SIA* capable non seulement de servir d'un locuteur dans un monologue mais aussi d'un partenaire d'interaction active échangeant les rôles de *locuteur* et d'*auditeur*. Pour être considéré comme un bon partenaire de communication, il doit posséder les compétences sociales nécessaires pour attirer et maintenir l'attention et l'implication de ses interlocuteurs. La capacité d'adaptation est une compétence sociale importante, innée chez l'humain, qui permet un tel engagement. Notre objectif est de créer des *SIA*s *engageantes* et *sociales* en les dotant d'une capacité d'*adaptation réciproque*. Malgré les diverses avancées des *SIA*s, plusieurs défis restent à relever pour créer des *SIA*s adaptatifs. Parmi les défis qui attendent d'être relevés, nous nous concentrons sur les aspects suivants.

Modélisation de l'adaptation

Lors d'une conversation, les humains communiquent avec leurs interlocuteurs par le biais d'une multitude de signaux *multimodaux* (verbaux et non verbaux) tout en adaptant constamment leur comportement à celui de leurs interlocuteurs. Tous les signaux provenant de soi-même et des partenaires d'interaction sont pris en compte pour la génération de son comportement. Pour que les *SIA*s puissent s'adapter de la même manière, il est essentiel qu'ils aient eux aussi la même capacité d'adaptation. L'*SIA* et l'interlocuteur humain doivent constamment s'adapter l'un à l'autre. Par conséquent, la modélisation de cette capacité d'*adaptation*, de la *multimodalité* et de la relation *interpersonnelle*, doit être effectuée lors du calcul des comportements des *SIA*s.

Évaluation de l'adaptation

L'évaluation des interactions a toujours été une tâche ardue. En particulier, l'évaluation de la présence d'une adaptation, et notamment sa quantification, n'est pas triviale. L'évaluation de l'adaptation serait très utile pour évaluer les interactions humain-humain et humain-agent.

Fonctionnalité en temps réel

Les *SIA*s sont conçues dans le but final d'être déployées pour de nombreuses applications réelles. Pour qu'elles puissent être présentées au public et démontrer leur utilité, elles doivent fonctionner en temps réel. Cet aspect temps réel doit être assuré pour tout système d'interaction car la communication se déroule généralement sans délai et est inattendue. Par conséquent, l'existence d'un retard pourrait nuire à l'expérience de l'utilisateur. Cela n'est pas souhaitable pour les *SIA*s qui recherchent l'engagement et la satisfaction de l'utilisateur.

Alignement temporel

Lors de la production d'un comportement, la parole et le geste correspondants sont alignés et synchronisés dans le temps. Cet aspect de l'alignement temporel est essentiel pour la génération d'un comportement. Il s'agit d'un problème difficile qui doit être résolu, en particulier pour assurer la *synchronisation temporelle* des signaux *multimodaux* pendant la perception et la génération en temps réel.

La portée de la thèse

Le thème principal de la thèse est la création d'un *SIA adaptatif* servant d'interlocuteur *social* et *engageant*. Notre objectif est d'activer la façade adaptative du *SIA* en lui fournissant une capacité d'*adaptation réciproque* et en garantissant son utilisation en *temps réel*. Nous relevons les défis susmentionnés en approfondissant les trois études de cette thèse. Chaque objectif d'étude et les questions de recherche associées sont détaillées.

Analyse des interactions humain-agent

Quel est le rôle de l'adaptation et comment la saisir ? Il est difficile d'obtenir une image claire de l'adaptation et de connaître son rôle exact dans la manière dont elle affecte d'autres aspects tels que l'engagement et les dimensions sociales. Afin de mieux comprendre l'adaptation et de découvrir comment elle peut être mesurée, nous analysons les interactions humain-humain dans le but d'appliquer ces connaissances aux interactions humain-agent.

Questions de recherche:

Dans cette étude, nous examinons la question de recherche suivante:

1. *Adaptation réciproque*: L'adaptation réciproque (synchronie et boucle d'entraînement) est-elle liée à l'engagement et/ou aux dimensions sociales de la chaleur et de la compétence des interlocuteurs ?

Génération de comportements avec capacité d'adaptation réciproque

Adopter un comportement approprié à une situation donnée, tel que des gestes cohérents avec la parole et correspondant au comportement des interlocuteurs, est une capacité inhérente qui donne l'impression d'être réalisée sans effort. Néanmoins, lorsque nous essayons de modéliser l'interaction entre les signaux *multimodaux* échangés, la modélisation n'est pas évidente. La relation entre les signaux sociaux (relations *interpersonnelles* et *intrapersonnelles*, et *multimodalité*) doit être prise en compte ainsi que leur *temporalité*. Cette étude vise à modéliser l'*adaptation réciproque* en capturant les *relations* et la *temporalité* des signaux échangés et à générer des comportements *SIA adaptatifs* (voir Chapitres 6 et 8). En particulier, nous restituons les gestes faciaux du *SIA* (expressions faciales et mouvements de la tête/du regard), car la richesse de l'expression faciale renforce les capacités de communication qui sont vitales pour la sociabilité des *SIA*s (Halberstadt [1983]).

Research Questions:

Dans cette étude, nous examinons les questions de recherche suivantes:

1. *Modélisation de l'adaptation réciproque*: Comment modéliser l'adaptation réciproque ?
2. *Impact de la modélisation de l'adaptation sur la dynamique interpersonnelle*: La dotation d'une capacité d'adaptation réciproque améliore-t-elle la dynamique interpersonnelle (en termes de synchronie et d'engagement) des comportements de *SIA* générés ?
3. *Impact de la modélisation de l'adaptation sur la qualité du comportement de SIA*: La dotation d'une capacité d'adaptation réciproque améliore-t-elle la qualité du comportement de *SIA* (en termes de la naturalité et de la vraisemblance humaine) ?
4. *Impact de la capture de la relation intrapersonnelle*: La modélisation explicite de la relation intrapersonnelle (modélisation de la relation entre les modalités) influence-t-elle la dynamique interpersonnelle et/ou la qualité du comportement du *SIA* ?

Système d'interaction humain-agent en temps réel

Pour étudier pleinement l'effet de la fourniture d'un *SIA* adaptatif, il est important que le *SIA* soit testé et évalué avec des utilisateurs humains réels dans un scénario de la vie réelle. L'impression du *SIA* peut être évaluée d'un point de vue à la troisième personne. Cependant, comme les *SIA*s visent à interagir avec des utilisateurs humains réels pour leur déploiement, il est préférable d'obtenir le retour

d'information des utilisateurs réels interagissant d'un point de vue à la première personne. Il est donc nécessaire de mettre en œuvre un système en *temps réel* et de l'utiliser pour vérifier l'importance de l'intégration de l'*adaptation réciproque* dans le *SIA* (voir Chapitre 7).

Questions de recherche:

Dans cette étude, nous nous penchons sur les questions de recherche suivantes:

1. *Effet du SIA adaptatif en temps réel sur l'expérience de l'utilisateur*: Un *SIA* adaptatif peut-elle améliorer l'expérience de l'utilisateur (perception de l'agent) ?
2. *Effet du SIA adaptatif en temps réel sur la performance de l'application*: Un *SIA* adaptatif peut-elle améliorer l'efficacité de la thérapie cognitivo-comportementale (TCC ou CBT en anglais; l'application santé que nous avons choisie) ?

Comportements non verbaux

Les signaux non verbaux, également appelés langage corporel, constituent une part importante des signaux de communication. Alors que la communication verbale transmet des informations par le biais d'un langage au contenu explicite, le comportement non verbal peut transmettre des informations de manière implicite et envoyer un message plus fort lorsqu'il est associé à un contexte verbal.

Le comportement non verbal est transmis par le "langage corporel", qui comprend les gestes, les expressions faciales, les mouvements du corps et le regard (Burgoon et al. [2011]). Nous considérons le comportement non verbal comme des signaux sociaux *multimodaux* qui transfèrent des informations de manière implicite ou explicite par le biais d'actions qui peuvent indiquer les attitudes ou les sentiments d'un individu sans utiliser de mots (c.-à-d. des informations lexicales). La prosodie (ou indices vocaux), comme la hauteur et le volume de la voix, est également un signal non verbal qui contient des informations pertinentes.

Lors de la transmission d'un message de communication, nous modifions, de manière intentionnelle ou non, notre comportement (Burgoon et al. [2011]). Ces intentions de communication sont généralement transmises par le biais de messages verbaux. Les comportements non verbaux véhiculent également de telles intentions, consciemment ou non. L'ajout de signaux non verbaux aux signaux verbaux peut transmettre le même message plus clairement et plus fermement à l'interlocuteur. En outre, la barrière de la langue,un problème inévitable pour la compréhension verbale et le retour d'information, peut être franchie grâce aux gestes. La compréhension de différentes langues n'est pas nécessaire pour manifester et reconnaître des sentiments et/ou des pensées. Le comportement non verbal est donc fondamental et influent sur la communication.

Les canaux non verbaux englobent les aspects comportementaux et les caractéristiques physiques qui constituent l'apparence physique des personnes (Knapp

et al. [2013]). Ils englobent les gestes, la posture, les expressions faciales, le comportement oculaire et le toucher. Ils comprennent également les caractéristiques de la parole, qui sont des indices vocaux.

Adaptation

Au cours d'une interaction, le comportement des interlocuteurs s'adapte. L'adaptation se fait en coordonnant (ou en synchronisant) le comportement de l'un avec celui de l'autre et en entraînant et en étant entraîné en permanence par le partenaire qui interagit. L'adaptation (e.g. la coordination ou la synchronisation du comportement) implique des phénomènes complexes tels que la perception de signaux sociaux et la réponse à ces signaux sociaux dans une fenêtre temporelle donnée (Chartrand and Lakin [2013], Burgoon et al. [1995]).

Les participants à la conversation échangent en réagissant aux signaux sociaux des autres. L'échange ne se fait pas simplement à tour de rôle entre les participants (avec un seul réacteur à la fois), mais la coordination implique différents processus tels que l'anticipation et la production de comportements. Condon and Ogston [1966] souligne qu'il existe des synergies intrapersonnelles qui se forment entre les comportements d'une personne et que ces synergies sont coordonnées entre les interlocuteurs (au niveau interpersonnel). Pour être coordonnés, ces comportements doivent correspondre les uns aux autres dans l'action et dans le temps (Hove and Risen [2009], Burgoon et al. [1995]). Pour la coordination interpersonnelle, il est essentiel que les comportements soient alignés au moment opportun (Delaherche et al. [2012]). Cette coordination des signaux sociaux peut également être appelée la *synchronie interpersonnelle*. Pickering and Garrod [2004] parle d'alignement défini comme l'adaptation des comportements verbaux des interlocuteurs. La coordination interpersonnelle des comportements est une opération continue qui se déroule automatiquement dans le temps au cours d'une interaction naturelle (Schmidt and Richardson [2008]). L'adaptation est donc dynamique.

Il est également important de noter que la coordination interpersonnelle, qui se fait passivement et involontairement pour s'adapter au comportement du partenaire, a un certain retard dans la perception et l'adaptation. Chartrand and Bargh [1999], qui affirment que la coordination interpersonnelle est causée par un comportement de mimétisme, appellent ce phénomène d'adaptation inconsciente (ou de mimétisme) l'effet caméléon. Cette perception du signal des interlocuteurs est sensible à l'alignement temporel. Pour les signaux non verbaux, l'alignement temporel (ou le délai de mimétisme) se situe dans une fenêtre temporelle de 2 à 4 secondes (Leander et al. [2012]).

Un entraînement continu se produit entre les interlocuteurs (Prepin and Pelachaud [2011]). Lorsqu'une personne adopte un comportement, elle entraîne le comportement mimétique de son interlocuteur. L'entraînement ne se limite pas à un simple mimétisme, mais il incite également l'émetteur du signal initial à continuer à adopter le même comportement ou à renvoyer le même signal. Nous appelons ce processus d'entraînement séquentiel une boucle d'entraînement.

Pour englober les divers aspects de l'adaptation, principalement la synchronie interpersonnelle et la boucle d'entraînement, nous désignons l'adaptation continue, dynamique et réciproque du comportement par le terme d'*adaptation réciproque*.

Pour le reste de notre travail, nous choisissons de définir les termes suivants comme suit:

***Synchronie* (ou *synchronie interpersonnelle*):** coordination interpersonnelle de signaux sociaux alignés au moment opportun, comme indiqué par Delaherche et al. [2012].

***Boucle d'entraînement*:** processus en boucle montré entre les interlocuteurs qui entraînent continuellement le comportement de mimétisme de leur interlocuteur, l'un après l'autre, comme mentionné par Prepin and Pelachaud [2011].

***Adaptation réciproque*:** adaptation du comportement des interlocuteurs au cours d'une interaction continue, dynamique et réciproque.

Interaction humain-agent (*HAI*)

L'interaction humain-agent (*HAI*) est l'interaction entre les humains et les *SIA*s. Son objectif est d'améliorer l'interaction entre l'humain et l'agent. Les progrès de l'*HAI* se concentrent sur une myriade d'aspects tels que l'engagement (Oertel et al. [2020]), la présence sociale (Pereira et al. [2014], Li [2015]) et le réalisme du comportement (Ferstl et al. [2021]) des agents. Dans le cadre de l'interaction humain-agent, de nombreux signaux sociaux de diverses modalités sont échangés. Comme dans l'interaction humain-humain, (l'humain et l'agent envoient et reçoivent des signaux *multimodaux* qu'ils interprètent et utilisent les informations perçues pour produire leur prochain comportement. La gestion de l'échange d'informations *multimodales* est également un aspect essentiel de l'interface humain-machine.

Etat de l'Art et Discussion

La génération de signaux non verbaux dépend du temps, comme les problèmes de séries temporelles. La rétention de mémoire présente dans les réseaux récurrents tels que RNN, LSTM et TCN s'est révélée très prometteuse pour la prévision des séries temporelles. Comme les comportements humains dépendent fortement des comportements antérieurs, cet aspect de la mémoire est également important pour notre situation. Le comportement devant être continu, il est de plus préférable d'utiliser la prédiction adaptative en ligne avec l'aspect de la prédiction basée sur les données temporelles précédentes d'une manière autorégressive.

Les modèles de pointe montrent comment la relation entre les signaux sociaux de soi-même (relation *intrapersonnelle*), les signaux des interlocuteurs (relation *interpersonnelle*) et les signaux multimodaux peut être modélisée. Pour notre travail, nous voulons modéliser l'*adaptation réciproque* en considérant les deux facettes de la *temporalité* (à la fois *intrapersonnelle* et *interpersonnelle*) et de la

multimodalité ainsi que l'aspect de la *continuité* pour la génération du comportement non verbal de notre agent. La modélisation de la multimodalité est absente dans Feng et al. [2017], Dermouche and Pelachaud [2019b] et la continuité n'est pas assurée pour Feng et al. [2017], Grafsgaard et al. [2018], Dermouche and Pelachaud [2019b], Jonell et al. [2020], Tuyen and Celiktutan [2022]. Bien que Ng et al. [2022] réponde à nos trois critères, il nécessite beaucoup de données d'entraînement. Dans notre cas, nous utilisons une petite base de données (réf. Chapitre 4), leur modèle n'est donc pas adapté à notre application. Nous proposons un nouveau modèle, le modèle *ASAP* (Augmented Self-Attention Pruning) présenté dans le Chapitre 6, qui rend les comportements non verbaux *continus* (pour le *locuteur* et l'*auditeur*) performants avec un petit ensemble de données. Il apprend également à capturer la relation *interpersonnelle* entre les interlocuteurs à partir des signaux *multimodaux* échangés afin de doter les *SIA*s d'une capacité d'*adaptation réciproque*. En outre, nous développons un autre modèle, le modèle *HI*^{2}*-ADAM* (Historical Intrapersonal Interpersonal ADAptive Multimodal) détaillé dans le Chapitre 8, qui capture également l'*adaptation réciproque* pour générer un comportement non verbal *adaptatif* et *continu* du *SIA* (pour les deux rôles) comme le modèle *ASAP*. Le modèle *HI*^{2}*-ADAM* intègre mieux l'adaptation entre les interlocuteurs en modélisant explicitement la relation *intrapersonnelle* avec les *historiques de modalité* (mémoire de modalité) et un encodage plus approfondi des signaux *multimodaux*.

Divers efforts ont été déployés pour quantifier la qualité des comportements non verbaux. Néanmoins, il n'existe pas encore de mesure parfaite pour les évaluer. En particulier, plusieurs aspects de la qualité du comportement tels que le *naturel* et la *ressemblance avec l'humain* peuvent être triviaux pour un humain, mais toujours très difficiles d'accès pour une machine (Fitrianie et al. [2020, 2021]). Ainsi, l'évaluation humaine reste une partie essentielle de l'évaluation du comportement (Feng et al. [2017], Karras et al. [2017], Chu et al. [2018], Sadoughi and Busso [2018], Alexanderson et al. [2020], Jonell et al. [2020], Yuan and Kitani [2020], Cai et al. [2021], Fitrianie et al. [2020]). Pour mieux évaluer la *synchronie interpersonnelle* entre l'humain et l'agent et pour compléter l'évaluation subjective, nous proposons l'utilisation de nouvelles mesures pour l'évaluation du comportement de l'agent, présentées au Chapitre 6, et de nouvelles *mesures d'adaptation réciproque* (mesures de synchronie et de boucle d'entraînement), introduites au Chapitre 5.

Corpus NoXi

Pour notre étude, nous utilisons la base de données NoXi (Cafaro et al. [2017]) qui est un corpus d'interactions en face-à-face médiées par l'écran. Elle contient des conversations dyadiques naturelles parlant d'un sujet commun. Chaque dyade d'interaction est composée d'une paire de participants ayant deux rôles différents, appelés expert et novice. L'expert est celui qui transfère des informations dans le but de partager ses connaissances sur un sujet et qui mène donc la conversation en parlant plus fréquemment et plus longtemps. Le novice (l'autre partenaire d'interaction) reçoit les informations et réagit aux propos de l'expert sur le sujet.

Le corpus NoXi se compose de 3 parties en fonction du lieu d'enregistrement (France, Allemagne et Royaume-Uni). Pour notre travail, nous n'utilisons que l'enregistrement du site français qui consiste en 21 interactions dyadiques réalisées par 28 participants avec une durée totale de $7h22min$.

Nous obtenons les caractéristiques du comportement non verbal des deux participants en interaction par extraction de caractéristiques. Pour chaque pas de temps, les caractéristiques visuelles et audio sont extraites en utilisant les outils OpenFace (Baltrušaitis et al. [2016]) et openSMILE (Eyben et al. [2010]) (après une phase de débruitage) respectivement et sont traitées séparément.

Pour analyser l'interaction humain-agent, nous ne nous contentons pas d'examiner les signaux de bas niveau (c.-à-. les caractéristiques extraites) échangés dans les interactions humain-humain du corpus NoXi, mais nous étudions également les signaux de haut niveau qui sont annotés. Les annotations de l'engagement (Dermouche and Pelachaud [2019a]) et des dimensions sociales (chaleur et compétence; Biancardi et al. [2017]) sont disponibles pour le corpus NoXi (disponibles avec l'outil d'annotation NOVA (Heimerl et al. [2019]). En outre, les annotations de l'état conversationnel sont récupérées automatiquement en effectuant une détection de l'activité vocale (VAD), qui est un classificateur binaire qui détecte la présence de la parole humaine dans l'audio, sur les fichiers audio déboisés.

Contribution

Cette thèse contribue aux communautés de recherche du *SIA* et du traitement des signaux *multimodaux* pour générer des comportements non verbaux *adaptatifs* du *SIA* en capturant les relations *intrapersonnelles* et *interpersonnelles* à partir de signaux *multimodaux*. Les contributions apportées par cette thèse sont les suivantes.

Proposition de nouvelles mesures d'adaptation réciproque

A partir de l'analyse de l'interaction humain-humain, nous avons étudié l'adaptation présente dans les conversations. Cette étude a servi de base pour proposer de nouvelles *mesures d'adaptation réciproque*. Les *mesures d'adaptation réciproque*, qui consistent en des mesures de synchronie et de boucle d'entraînement, ont été utilisées pour étudier la relation entre l'adaptation et les dimensions de l'engagement, de la chaleur et de la compétence. Les nouvelles mesures proposées ont montré leur utilité dans l'évaluation de la qualité de l'interaction humain-agent. Elles ont été utilisées pour les évaluations objectives de notre système *IAVA* (réf. Chapitre 7) et du modèle *HI*^{2}*-ADAM* (réf. Chapitre 8).

Rendre les comportements adaptatifs des *SIA*s

La capacité d'*adaptation réciproque*, qui est une capacité importante innée chez les humains pour les communications *interactives* et *engageantes*, est conférée aux *SIA*s en modélisant la *multimodalité*, la relation *interpersonnelle* et/ou la relation *intrapersonnelle*. Nous générons des comportements *adaptatifs* des agent virtuels

via notre modèle *ASAP* (réf. Chapitre 6) et le modèle *HI^2-ADAM* (réf. Chapitre 8). Il a été démontré que le comportement rendu par *SIA* surpasse les techniques de pointe en générant un comportement *naturel*, *humain*, *synchronisé* et *engageant*. Grâce à nos *mesures d'adaptation réciproque* (réf. Chapitre 5), nous avons également pu valider objectivement que les comportements prédits étaient effectivement *réciproquement adaptatifs* ainsi que l"utilité de ces mesures dans l'évaluation de la qualité de l'interaction entre l'humain et l'agent.

Développement d'un système *SIA* interactif et adaptatif en temps réel

Le but ultime du développement d'agents animés, qu'il s'agisse de *SIA*s ou de robots, est de les déployer en *temps réel* avec l'utilisateur final humain. Le fonctionnement en *temps réel* est essentiel, en particulier pour l'adaptation dans l'interaction humùain-agent ou humain-robot. Nous avons créé un système de *SIA interactif* et *adaptatif*, notre système *IAVA* (réf. Chapitre 7), qui garantit l'aspect *temps réel*. En appliquant le système *IAVA* à l'application médicale du CBT, nous avons vérifié l'efficacité des *SIA*s avec des capacités d'*adaptation réciproque* pour donner une impression positive aux utilisateurs (être perçu comme *naturel*, *humain*, *engageant*, *synchrone*, et *établir un rapport*) et pour améliorer l'effet du CBT. En outre, pour démontrer la possibilité d'utiliser notre système de *SIA* adaptatif dans d'autres applications, nous avons également testé notre système pour la SST. Nous avons constaté que les utilisateurs ont une impression similaire de l'agent pour les deux applications, SST et CBT, malgré la nature différente des scénarios. En outre, nous avons collecté une base de données d'interactions humain-agent (*CBT-HAI DB*). Les interactions CBT entre le *SIA* et l'utilisateur ont été enregistrées et la base de données a été mise à la disposition de la communauté des chercheurs (après signature du formulaire EULA).

Mots-clés: Interactions Humain-Agent, Adaptation Réciproque, Agents Conversationnels Animés, Génération des Expressions Faciales, Apprentissage profond, Multimodalité

Abstract

Information is transferred from one person to another via communication. Through this transfer, we convey our thoughts and intentions via *multimodal* signals such as words, gestures, and prosody. This exchange of signals is a two-way process of sending and receiving where the behaviors of the interlocutors adapt to each other. Such adaptation is continuous, dynamic, and reciprocal which we refer to as *reciprocal adaptation*. Adapting to others allows interactions to be engaging and effective. Endowing such capacity to Socially Interactive Agents (*SIA*s), physical or virtual embodied agents (such as virtual agents and robots), can make them more *social* and *engaging*, and perceived as *natural* and *human-like*. Nevertheless, this endowment is a challenging task. The agent needs to know how to adapt as both a *speaker* and a *listener* while emitting behaviors related to its own *speech* synchronized over its modalities, *intrapersonal* relationship, and with its interlocutor's behaviors, *interpersonal* relationship. The central focus of this thesis is to develop an adaptive *SIA* with *reciprocal adaptation* capabilities. We propose computational models, *ASAP* and *HI*^{2}*-ADAM*, to render *SIA*'s adaptive behaviors as both a *speaker* and a *listener*. *ASAP* generates *adaptive* and *continuous* behavior using *multimodal* signal information from its user and itself by modeling the *interpersonal* relationship between them. *HI*^{2}*-ADAM* captures the *reciprocal adaptation* and *intrapersonal* relationship in an explicit way by modeling the *modality history* of each interlocutor and learning from the relation between these different *histories*. As it is important for agents to act as interactive partners and continuously adapt their behaviors in *real time*, we create a *real-time interactive* and *adaptive* agent, *IAVA* system, and provide new measures, *reciprocal adaptation measures*, for the evaluation of human-agent interaction quality.

Keywords: Human-Agent Interaction, Reciprocal Adaptation, Socially Interactive Agents, Facial Expression Generation, Deep Learning, Multimodality

Acknowledgment

I would like to thank everyone for their presence, help, and contribution. This thesis would not have existed without them.

First and foremost, I would like to express my deepest gratitude to my supervisors Catherine Achard and Catherine Pelachaud for your support and guidance throughout this three-year PhD journey. It was an honor to work with you and I was able to gain new insights and understanding thanks to your advice. It was an unforgettable experience and I am greatly thankful to both of you for everything.

I would also like to thank the entire member at Institut des Systèmes Intelligents et de Robotique (ISIR). In particular, I would like to thank my colleagues Mireille Fares, Liu Yang, Michele Grimaldi, Lucie Galland, Fabien Boucaud, Nezih Younsi, Fajrian Yunus, Sooraj Krishna, and Silvia Tulli. You added some fun to this long journey and I was able to get various inspirations and motivations from you.

Many thanks to my precious friends. I am so lucky to have met you and have you by my side. Special thanks to my childhood friends Connie Shin, Iris Lee, and Younga Kang and to my close friends Hyeju Park and Hyunmin Lee who helped me get through hard times.

Lastly, my biggest thanks go to my parents, my mother Young Hee Lee and my father Sanghyun Woo, and my brother Alex Woo who were always there for me supporting every step of my life.

Contents

List of Figures

List of Tables

Nomenclature

General

$(,)$	pair
$(.)$	set
$[.]$	concatenation
$\neg$	not
F	f-value of ANOVA
p	p-value for statistical significance (ANOVA, Tukey's HSD, two-tailed t-test)

Math Operators

$\|x\|$	absolute value of x
Δ	delta
$\max(\cdot)$	maximum
Σ	summation
$\sigma(x)$	sigmoid function
$\sqrt{\cdot}$	square root
$\widehat{X}$	estimate of X
$CA(\cdot)$	cross-attention
$SA(\cdot)$	self-attention
X	matrix
x	scalar
x^n	nth power

Features

AU	action unit
$AU1$	inner brow raiser
$AU12$	lip corner puller for smile
$AU2$	outer brow raiser
$AU4$	brow lowerer
$AU5$	upper lid raiser
$AU6$	cheek raiser
$AU7$	lid tightener
$F0$	fundamental frequency
G_x	eye movement along x axis
G_y	eye movement along y axis
R_x	head rotation along x axis
R_y	head rotation along y axis
R_z	head rotation along z axis
$t0_{ts}$	onset point of the listener's voice activity for the identification of turn-shift and backchannel moments

Networks Components, Layers, and Units

$\widehat{Y}$	generated *SIA* behavior
c	cell size
D	dense layer
d	depth of MHA
E	encoder
h	number of attention heads of MHA (head size)
h'	number of selected attention heads of MHA for the custom pruning mask
K	key of attention mechanisms
K	value of attention mechanisms
Q	query of attention mechanisms
T	input sequence time-step length

t	current time-step
$t+1$	next time-step
Y	real (ground truth) *SIA* behavior
Z	embedding

Chapter 1

Introduction

Contents

This chapter introduces the field of Socially Interactive Agents (*SIAs*) and presents the research context and aim upon which the thesis is built. The thesis objectives and research questions are presented and the contributions and publications are briefly listed. It is finished with an outline of the thesis structure.

1.1 Socially Interactive Agents

Socially Interactive Agents (*SIA*s; Lugrin et al. [2021]) are embodied agents that are physical or virtual, such as robots or virtual agents (see Figure 1.1), capable of autonomously carrying out natural conversations with people (i.e. their users) by exchanging *multimodal*, verbal and nonverbal, signals (Ball et al. [2000]) in a socially intelligent manner. The field of *SIA* has grown for more than 20 years under manifold names such as Intelligent Virtual Agents, Embodied Conversational Agents, and Social Robotics (Cassell [2001], Dautenhahn [1998]). They all share a similar definition pursuing the same purpose of advancing the research and development of socially intelligent agents displaying behaviors autonomously through physical or virtual embodiments.

*SIA*s transmit their message (intention or feeling) verbally via words and emit *human-like* nonverbal behavior, such as *facial*, *body*, and *hand* gesturing during *speech* (Cassell et al. [2000]), to complement their message.

Figure 1.1 Illustration of *SIA*s. From left to right: Greta (Niewiadomski et al. [2009]), Meta avatar [a], Furhat robot (Al Moubayed et al. [2013]), Miroki robot [b], and Pepper robot (Pandey and Gelin [2018]).

[a] https://developer.oculus.com/documentation/unity/meta-avatars-overview/
[b] https://enchanted.tools/robot

Embodied agents have the common goal of improving communication and the user's experience (engagement, rapport, and liking). Several works have focused on enhancing the interaction itself for a smoother experience with the agent (Bailenson and Yee [2005], Huang et al. [2010], Ritschel et al. [2017], Weber et al. [2018]).

The use of *SIA*s has been seen for various applications showing their usefulness in teaching/coaching (Anderson et al. [2013], Pecune et al. [2016], Jones and Castellano [2018], Pereira Santos et al. [2023]), assisting (Biancardi et al. [2021], Sidner et al. [2018]), and providing healthcare (Raffard et al. [2018], Ring et al. [2016], Shidara et al. [2022], Khamis et al. [2021]). Especially for

medical treatment, the employment of *SIA*s has greatly increased in recent years proving the utility of *SIA*s (Khamis et al. [2021], Shidara et al. [2022].

1.1.1 Research context

The goal of *SIA* research is to create *SIA*s that are capable of interacting with people in a *social* and *engaging* way. Ensuring such aspects of *socialness*, *engagement* (starting, maintaining, and ending the perceived connections to each other during an interaction; Sidner et al. [2003]), and *rapport* (positive affect, mutual attention, and coordination; Tickle-Degnen and Rosenthal [1990]) is essential to grab the attention of its users and continuously carry out the interaction. To render effective communication, the agents taking the role of interacting partners must play active roles of both *speaker* and *listener*.

Human-human interaction studies have shown that communication is *multimodal*. Different channels (or modalities composed of verbal and nonverbal behaviors) of communication participate in passing one's intention and emotions. They are synchronized both *intrapersonally* (between the cues of the same person) and *interpersonally* (between interlocutors) illustrated in Figure 1.2.

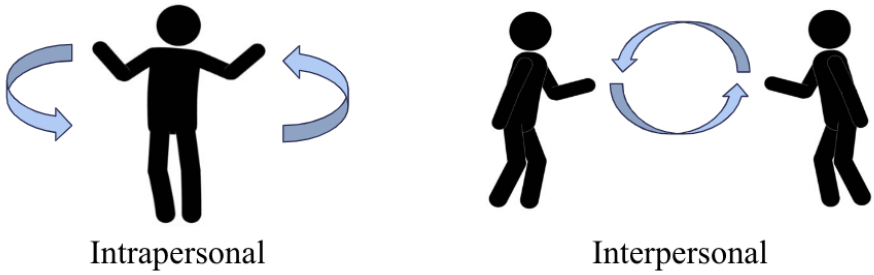


Figure 1.2 Illustration of intrapersonal and interpersonal relationships.

For *SIA*s to act as interlocutors, they need to not only pass their message by aligning their *multimodal* signals derived from themselves (considering the *intrapersonal* relationship; Knapp et al. [2013]) but also constantly and reciprocally adapt to their interlocutors by coordinating their behaviors to their interlocutors' multimodal signals (taking into account the *interpersonal* relationship; Burgoon et al. [1995]). The display of such adaptation capability is key to *interpersonal* relationships (Cappella [1991]) and may enable *SIA*s to be perceived as *social*, *engaging*, and establishing *rapport* (Biancardi et al. [2021], Oertel et al. [2020], Delaherche and Chetouani [2010], Gupta et al. [2019], Huang et al. [2010], Raffard et al. [2018]).

1.1.2 Research aim

*SIA*s have the goal to not solely serve as a speaker in a monologue but also as an active interacting partner exchanging the speaking roles of *speaker* and *listener*. For them to communicate with their interlocutors, they must have the social skills to attract and maintain their interlocutors' attention and involvement. One important social skill, innate to humans, that enables such *engagement* is the adaptation

capacity that is *continuous*, *dynamic*, and *reciprocal*. We refer to such adaptation as *reciprocal adaptation* of which its endowment may allow *SIA*s to be considered as a good communication partner. Despite the diverse advances of *SIA*s, various challenges remain to be solved for the creation of adaptive *SIA*s. The main challenges that await to be addressed are as follows.

Figure 1.3 Illustration of Human-Agent Interaction with reciprocal adaptation.

Adaptation modeling:

In a conversation, humans communicate with their interlocutors via a multitude of *multimodal* signals (verbal and nonverbal) while constantly adapting their behavior to those of their interlocutors. All signals originating from oneself and interacting partners are taken into account for the generation of one's behavior. For *SIA*s to adapt in the same way, it is essential for them to also have the same adaptation capability. The *SIA* and human interlocutor should constantly adapt to each other as shown in Figure 1.3. Hence, the modeling of such adaptation skill, of *multimodality* and *interpersonal* relationship, must be performed when computing *SIA* behaviors.

Adaptation assessment:

Interaction evaluation has always been an onerous task. Notably, the assessment of the presence of adaptation, particularly quantifying it, is not trivial. The assessment of adaptation would be substantially helpful for evaluating human-human and human-agent interactions.

Real-time functionality:

*SIA*s are designed with the final objective of being deployed for copious real-life applications. For them to be shown to the public and demonstrate their helpfulness, they need to function in *real time*. This *real-time* facet needs to be assured for any interaction system as communication generally proceeds with no delay and is unexpected. Thus, the existence of a delay might deter the user experience which is unwanted for *SIA*s that seek user *engagement* and satisfaction.

Temporal alignment:

When producing a behavior, the corresponding speech and gesture are *timely aligned* and *synchronized*. This aspect of *time alignment* is critical for behavior generation. It is a challenging problem that must be solved especially assuring the *temporal sync* of *multimodal* signals during perception and generation in *real time*.

1.2 Thesis Scope

The major theme of the thesis is to create an *adaptive SIA* serving as a *social* and *engaging* interlocutor. Our goal is to enable the *adaptive* facade of *SIA* by supplying the *SIA* with *reciprocal adaptation* capacity and ensuring its *real-time* use. We address the aforementioned challenges by digging deeper into the three studies of this thesis. Each study objective and associated research questions are detailed.

1.2.1 Human-Agent interaction analysis

What is the role of adaptation and how can adaptation be captured? It is hard to get a clear image of adaptation and know its exact role in how it affects other aspects such as engagement and social dimensions. With the aim of getting a better understanding of adaptation and finding out how it can be measured, we analyze human-human interactions with the goal of applying this knowledge to human-agent interactions (see Chapter 5).

Research Questions:

In this study, we investigate the following research question:

1. *Reciprocal adaptation*: Is *reciprocal adaptation* (synchrony and entrainment loop) related to engagement and/or social dimensions of warmth and competence of the interlocutors?

1.2.2 SIA behavior generation with reciprocal adaptation capacity

Exhibiting appropriate behavior for a given situation, such as gestures that are coherent with speech and matching the interlocutors' behavior, is an inherent ability that makes it seem to be done effortlessly. Nevertheless, when we try to model the interplay between the exchanged *multimodal* signals, the modeling is not evident. The relation between social signals (*interpersonal* and *intrapersonal* relationship, and *multimodality*) needs to be considered as well as their *temporality*. This study aims to modelize the *reciprocal adaptation* by capturing the *relations* and *temporality* of interchanged signals and to generate *adaptive SIA* behaviors (see Chapters 6 and 8). In particular, we render the *SIA*'s *facial gestures* (facial expressions and head/gaze movements) as the richness of facial expression elevates communication skills which are vital for the sociability of *SIA*s (Halberstadt [1983]).

Research Questions:

In this study, we investigate the following research questions:

1. *Reciprocal adaptation modeling*: How can we model the *reciprocal adaptation*?
2. *Impact of adaptation modeling on interpersonal dynamics*: Does the endowment of *reciprocal adaptation* capability improve *interpersonal dynamics* (in terms of synchrony and engagement) of the generated *SIA* behaviors?
3. *Impact of adaptation modeling on SIA behavior quality*: Does the endowment of *reciprocal adaptation* capability improve the *quality* of the *SIA*'s behavior (in terms of naturalness and human-likeliness)?
4. *Impact of capturing intrapersonal relationship*: Does the explicit modeling of the *intrapersonal relationship* (modelization of the relation between *modality*) influence the *interpersonal dynamics* and/or the *quality* of the *SIA*'s behavior?

1.2.3 Real-time system of human-agent interaction

To fully investigate the effect of providing an *adaptive SIA*, it is important for the *SIA* to be tested and evaluated with real human users in a real-life scenario. The impression of the *SIA* may be assessed via a third-person point of view. However, as *SIA*s aim to interact with real human users for their deployment, it is better to get the feedback of the actual users interacting in a first-person point of view. Therefore, the implementation of a *real-time* system is necessary and the system should be used to verify the significance of the embedding of *reciprocal adaptation* to the *SIA* (see Chapter 7).

Research Questions:

In this study, we investigate the following research questions:

1. *Effect of real-time adaptive SIA on user experience*: Can an *adaptive SIA* enhance the user experience (agent perception)?
2. *Effect of real-time adaptive SIA on application performance*: Can an *adaptive SIA* improve the effectiveness of CBT (user mood and state change; our chosen healthcare application)?

1.3 Thesis contribution

The main focus of this thesis is to develop an *interactive* and *adaptive SIA*. We develop a *real-time adaptive SIA* of which its nonverbal behaviors, notably its facial expressions and head/gaze movements, are generated by capturing *intrapersonal* and/or *interpersonal* relationships from *multimodal* signals. The contributions of this thesis are discussed below.

Novel measures of reciprocal adaptation

Our first contribution is the proposition of novel *measures of reciprocal adaptation*. The *adaptation measures* are introduced and explained in detail in Chapter 5. The relation between adaptation and the dimensions of engagement, warmth, and competence was checked via our newly proposed measures. Moreover, the usefulness of the measures was verified for the assessment of human-agent interaction quality. The *adaptation measures* were used for the objective evaluations of our *IAVA* system (ref. Chapter 7) and *HI^2-ADAM* model (ref. Chapter 8).

Parts of Chapter 5 appeared in ICAART (Woo et al. [2023e]).

Adaptive SIA behavior generation models

Reciprocal adaptation skill is a crucial one that plays a fundamental role in human communication enabling *interactive* and *engaging* interactions. *SIA*s can acquire this ability by modeling the interaction aspects of *multimodal* communication, and *interpersonal* and *intrapersonal* dynamics. We propose two different computational models to endow the *reciprocal adaption* capability which are:

- *ASAP* model (ref. Chapter 6) modeling the *reciprocal adaptation* focusing on the aspects of *interpersonal temporality* (via self-attention pruning technique), *multimodality* (multimodal signal encoding), and behavior prediction *continuity* (via autoregressive adaptive online prediction technique).
- *HI^2-ADAM* model (ref. Chapter 8) generating *adaptive SIA* behavior as both *speaker* and *listener* by encoding the *multimodality*, and *intrapersonal* (explicit modeling of *modality histories*) and *interpersonal* relationships.

The generated *SIA* behavior, of both models, outperforms the state-of-the-art methodologies in terms of *naturalness*, *human-likeness*, *synchrony*, and *engagement*.

Parts of Chapter 6 appeared in 28th International Conference on Intelligent User Interfaces (Woo et al. [2023d]) and parts of Chapter 8 appeared in arXiv preprint (Woo et al. [2023a]).

Real-time adaptive SIA

Real-time functionality is important for any system interacting with human end-users. To validate the usefulness of *SIA*s with *reciprocal adaptation* ability, we developed a *real-time adaptive SIA* system, our *IAVA* system (ref. Chapter 7). The efficiency of *adaptive SIA*s is shown for the applications of Cognitive Behavior Therapy (CBT) and Social Skills Training (SST). The system renders a positive impression to its users, as it is perceived to be *natural*, *human-like*, *engaging*, *in sync*, and building a *rapport*. Furthermore, it proved its serviceability in improving the CBT effect. A new human-agent interaction database (*CBT-HAI DB*) has been collected of the CBT interactions between the *real-time adaptive SIA* and human users which is available to the research community upon demand.

Parts of Chapter 7 appeared in ACM International Conference on Intelligent Virtual Agents (IVA '23) (Woo et al. [2023c,b]) and in 2023 International Conference on Multimodal Interaction (Saga et al. [2023b]), and submitted to IJHCS (Woo et al. [2023f]).

1.4 Publications and Submissions

- Jieyeon Woo, Catherine Pelachaud, and Catherine Achard. Asap: Endowing adaptation capability to agent in human-agent interaction. In *28th International Conference on Intelligent User Interfaces*, 2023d
- Jieyeon Woo, Catherine Pelachaud, and Catherine Achard. Reciprocal adaptation measures for human-agent interaction evaluation. In *ICAART*, 2023e
- Jieyeon Woo, Liu Yang, Catherine Pelachaud, and Catherine Achard. Is turn-shift distinguishable with synchrony? In *International Conference on Human-Computer Interaction*, pages 419–432. Springer, 2023h
- Jieyeon Woo, Liu Yang, Catherine Achard, and Catherine Pelachaud. Are we in sync during turn switch? In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–4. IEEE, 2023g
- Jieyeon Woo. Development of an interactive human/agent loop using multimodal recurrent neural networks. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 822–826, 2021
- Jieyeon Woo, Catherine Pelachaud, and Catherine Achard. Creating an interactive human/agent loop using multimodal recurrent neural networks. In *WACAI 2021*, 2021
- Jennifer Hamet Bagnou, Elise Prigent, Jean-Claude Martin, Jieyeon Woo, Liu Yang, Catherine Achard, Catherine Pelachaud, and Céline Clavel. A framework for the assessment and training of collaborative problem-solving social skills. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pages 381–384, 2021
- Jieyeon Woo, Michele Grimaldi, Catherine Pelachaud, and Catherine Achard. Iava: Interactive and adaptive virtual agent. In *ACM International Conference on Intelligent Virtual Agents (IVA '23)*, 2023c
- Jieyeon Woo, Michele Grimaldi, Catherine Pelachaud, and Catherine Achard. Conducting cognitive behavioral therapy with an adaptive virtual agent. In *ACM International Conference on Intelligent Virtual Agents (IVA '23)*, 2023b
- Takeshi Saga, Jieyeon Woo, Alexis Gerard, Hiroki Tanaka, Catherine Achard, Satoshi Nakamura, and Catherine Pelachaud. An adaptive virtual agent platform for automated social skills training. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2023b

- *(Preprint)* - Jieyeon Woo, Mireille Fares, Catherine Pelachaud, and Catherine Achard. Amii: Adaptive multimodal inter-personal and intra-personal model for adapted behavior synthesis. *arXiv preprint arXiv:2305.11310*, 2023a
- *(Under revision, Submitted to IJHCS)* - Jieyeon Woo, Kazuhiro Shidara, Catherine Achard, Hiroki Tanaka, Satoshi Nakamura, and Catherine Pelachaud. Adaptive virtual agent: Design and evaluation for real-time human-agent interaction. *International Journal of Human-Computer Studies*, 2023f

1.5 Thesis Outline

The thesis is organized into 5 parts which are as follows.

Part I presents the theoretical background, discusses related work, and introduces the corpus. More specifically, Chapter 2 establishes the background knowledge around *SIA*, human-agent interaction, nonverbal behavior, and adaptation. Chapter 3 provides insight into existing approaches of sequence prediction, nonverbal behavior generation, multimodal signal processing, and human-agent interaction evaluation. Chapter 4 explains the corpus.

Part II focuses on human-human interaction analysis. Analysis of human-human interactions around synchrony is done to propose new *reciprocal adaptation measures* in Chapter 5.

Part III is devoted to the behavior generation of *SIA* with *reciprocal adaptation*. In Chapter 6, we present our *ASAP* model, a model that models the *interpersonal relationship* (or *reciprocal adaptation*) rendering *social* and *engaging SIA* behaviors.

Part IV describes the development and evaluation of our *real-time adaptive SIA* system. In Chapter 7, we detail the system architecture design and evaluate with Cognitive Behavioral Therapy (CBT) and Social Skills Training (SST) as its proof-of-concepts.

Part V demonstrates the modeling of *reciprocal adaptation* with *intrapersonal modality history*. We propose our *HI^2-ADAM* model, Chapter 8, which captures the *reciprocal adaptation* and *intrapersonal relationship* in an explicit way by modeling each *modality history* (or *memory*) of each interlocutor and learning from the relation between these different *histories*.

Part I

Background, Related Work, and Corpus

Chapter 2

Background

Contents

The interaction between *SIA*s and humans (Human-Agent Interaction or *HAI*) is an important field that needs to be studied to ameliorate the use of embodied agents. Like human-human interaction, *HAI* also consists of social signals, verbal and nonverbal, exchanged between interlocutors, and the adaptation is present throughout the interaction. In this chapter, we draw attention to the theoretical background and definition of nonverbal behaviors, adaptation, and *HAI* to present the founding concept of our work.

2.1 Nonverbal behaviors

Nonverbal signals, which are also referred to as body language constitute a major part of communication signals. While verbal communication transfers information through language containing explicit content, nonverbal behavior can convey information implicitly and send a stronger message when combined with verbal context.

2.1.1 Terminology of nonverbal behavior

Nonverbal behavior is conveyed through "body language" including gestures, facial expressions, body movements, and gaze (Burgoon et al. [2011]). We consider nonverbal behavior as *multimodal* social signals that transfer information implicitly or explicitly via actions that can indicate an individual's attitudes or feelings without using words (i.e. lexical information). Prosody (or vocal cues), such as pitch and loudness of voice, is also a nonverbal signal that contains pertinent information.

2.1.2 Importance in communication

When transmitting a communication message, we intentionally or unintentionally vary our behavior (Burgoon et al. [2011]). Such communicative intentions are generally transferred via verbal messages. Nonverbal behaviors also covey such intent consciously and unconsciously. Adding nonverbal signals to verbal cues can transmit the same message more clearly and strongly to the interlocutor. Furthermore, the language barrier (an inevitable problem for verbal comprehension and feedback) can be broken via gestures. The understanding of different languages is not needed to manifest and recognize feelings and/or thoughts. Nonverbal behavior is fundamental and influential in communication.

2.1.3 Types

Nonverbal channels encompass the behavioral aspects with physical characteristics that make up people's physical appearance (Knapp et al. [2013]). These embrace gestures, posture, facial expressions, eye behavior, and touch. They also include speech characteristics which are vocal cues.

Gestures:

Gestures are body movements made up of a combination of head, arm, and hand motions. They are frequently studied in two groups of gestures that are speech-independent and those that are speech-related.

- Speech-independent gestures do not accommodate verbal cues. Nevertheless, they do transfer a direct meaning (via linguistic wordings or phrases produced at the same time) which can be considered as a direct translation of the verbal message. Such translated body signals, which are often culture-specific (Kita [2009]), are signals that people abided by tacit agreement.

They are represented by signs such as the head nod for agreement or the hand sign of "V" for victory that are solely produced without verbal signals.

- Speech-related gestures are deeply linked with or assist speech. Such types of gestures generally have the purpose of illustrating the content of what is being said. These motions help the interlocutor to understand better the message that is being transmitted. They can visually picture the information, detect which word or phrase is being emphasized by the speaker, infer the process or path of thought, or directly see which object or location is being pointed at.

Posture:

The posture is the position of the body (in particular the torso). Its inclination indicates the degree of engagement, attention, or involvement within an interaction. The stance of leaning forward can be interpreted as a positive sign of engagement and closeness while leaning back will be perceived as being bored or keeping a distance from the other interacting partner (D'Mello et al. [2007]). The mimicry or mirroring of posture may also reflect rapport (Sharpley et al. [2001]).

Facial expressions:

Facial expressions are critical in comprehending people's actions and behavior. They are the most representative cues for expressing emotional state and attitude (Ekman and Friesen [1978], Argyle [2013]). Daily people make various faces conveying emotions such as anger, disgust, fear, sadness, excitement, boredom, sympathy, calmness, awkwardness, and confusion. By displaying such expressions they provide feedback on the signal sender's state. With the variation and enrichment of the facial expressions the interaction flow is managed and the communication skills of the exhibitor are heightened (Halberstadt [1983]).

Eye behavior:

Eye behavior is the eye movement of where we are looking at. When observing, not only does the point where we look matter but the duration (how long we look at or away) and the timing (when we look) also carry weight in communication. Especially during a conversation, the mutual gaze (interlocutors looking into each others' eyes; or eye contact) is an essential factor that serves as an indication of social presence, interest, engagement, or impression (Mason et al. [2005], Kompatsiari et al. [2017]).

Touch:

Touch is another way of communicating especially affect. It can convey an extensive range of meanings. Jones and Yarbrough [1985] identified distinct and relatively unambiguous meanings of social touch which are support, appreciation, inclusion, sexual interest/intent, affection, playful affection, playful aggression, compliance, attention-getting, announcing a response, greetings, and depar-

ture. Touch is heavily dependent on the relationship between the touchee and the toucher. It is effective for social bonding and disclosing emotions (Chatel-Goldman et al. [2014], Hertenstein et al. [2009]).

Vocal cues:

Vocal cues (or speech prosody) are vocal expressions that convey meaning or information about the speaker's emotions, intents, or attitudes. Prosody is similar to facial expressions in the way that it is affected by the surrounding social components (Scherer et al. [1991], Argyle [2013]). Vocal cues are vocalization characteristics represented by various frequencies and intensities. We can change the way we speak (tone of voice) through the choice of pitch (high or low), volume (loud or soft; or loudness), speed (fast or slow; or pace of speaking or speech rate). Moreover, the same phrase can be spoken differently, stressing words or phrases, by changing the pronunciation, enunciation, or articulation. The rhythm can also be modified by inserting break points with pauses or adding musicality. These acoustic features hint at the state of the speaker containing pragmatic, synthetic, emotional, and contextual information. In addition, depending on the prosodic cues, the same utterance could be interpreted differently. For example, the word "ok" could be understood as a positive or negative response depending on how it is said.

2.2 Adaptation

During an interaction, behavior adaptation between interlocutors takes place. The adaptation is done by coordinating (or synchronizing) one's behavior to that of the other and by constantly entraining and being entrained by the interacting partner. Adaptation (e.g. behavior coordination or synchronization) involves complex phenomena such as perceiving social signals and responding to these social signals within a given time window (Chartrand and Lakin [2013], Burgoon et al. [1995]).

2.2.1 Reciprocal Adaptation

Conversation participants exchange by reacting to each other's social signals. The exchange is not simply alternated by taking turns between the participants (having a single reactor at the time), but the coordination involves different processes such as anticipating and producing behaviors. Condon and Ogston [1966] point out that there are intrapersonal synergies that are formed between one's behaviors and these synergies are coordinated across the interlocutors. They split the coordination into two types: *intrapersonal coordination* for the behavior coordination within oneself and *interpersonal coordination* for behavior coordination between multiple people in an interaction. To be coordinated these behaviors should match each other in action and time (Hove and Risen [2009], Burgoon et al. [1995]). We can note that Chartrand and Lakin [2013] used the term behavioral mimicry (a type of adaptation) when referring to the display of the same behavior at the same time by 2 or more participants. For interpersonal coordination, an essential aspect

is that the behaviors are timely aligned (Delaherche et al. [2012]). This coordination of social signals may also be referred to as *interpersonal synchrony*. Pickering and Garrod [2004] talk about alignment defined as the adaptation of interlocutors' verbal behaviors. The interpersonal coordination of behaviors is an ongoing operation that turns automatically in time during a natural interaction (Schmidt and Richardson [2008]). Thus, adaptation is dynamic.

It is also important to note that interpersonal coordination, which is done passively and unintentionally to match the interacting partner's behavior, has a certain delay in perception and adaptation. Chartrand and Bargh [1999], who state that interpersonal coordination is caused by mimicry behavior, call this unconscious adaptation (or mimicry) phenomenon the chameleon effect. This perception of interlocutors' signal is sensible to temporal alignment. For nonverbal signals, the temporal alignment (or the mimicry time delay) is along a time window of 2 to 4 seconds (Leander et al. [2012]).

Continuous entrainment occurs between the interlocutors (Prepin and Pelachaud [2011]). When a person shows a behavior, it entrains the mimicry behavior of their interactant. The entrainment doesn't end with a simple mimicry but it also rentrains the initial signal sender to continue performing the same behavior or to resend the same signal. We refer to this process of sequential entrainment as *entrainment loop*.

Adaptation, notably synchrony and entrainment, is closely linked to interaction (e.g. human-human, human-machine, and human-agent interactions) dimensions. The dynamic mutual adaptation has been shown to boost the engagement level of the interlocutor and to build a stronger rapport between the interlocutors within the interaction Delaherche et al. [2012], Raffard et al. [2018].

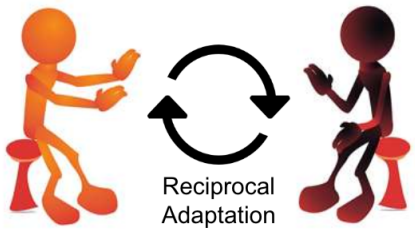


Figure 2.1 Illustration of reciprocal adaptation.

To encompass the diverse adaptation aspects, mainly interpersonal synchrony and entrainment loop, we refer to the continuous, dynamic, and reciprocal behavior adaptation as *reciprocal adaptation* (illustrated in Figure 2.1).

2.2.2 Definitions of relevant terms

For the rest of our work, we choose to define the following terms as:

***Synchrony* (or *interpersonal synchrony*):** interpersonal coordination of social signals that are timely aligned, as stated by Delaherche et al. [2012].

***Entrainment loop*:** looped process shown between interlocutors of continuously entraining the mimicry behavior of their interactant one after another as mentioned by Prepin and Pelachaud [2011].

***Reciprocal adaptation*:** behavior adaptation of interlocutors during an interaction that is continuous, dynamic, and reciprocal.

2.3 Human-Agent Interaction (HAI)

The domain of Human-Agent Interaction (*HAI*) derives from the research on Human-Machine Interaction (*HMI*; Human-Computer Interaction or Man-Machine Interaction Boy [2017], Dix [2003]) which studies how humans interact with machines. To get a better picture of what *HAI* is, we need to go back to the origins of *HMI* and see how it started. With the appearance of machines, their advantage of automating manual work has raised the discussion on the difference in the feasibility of tasks. The comparison between humans and machines in whether the same assignment could be done and which performs better in terms of perfection, speed, and cost-effectiveness became the interest of researchers. Fitts [1951], one of the pioneers of *HMI*, attempted to distinguish humans and machines by systematically characterizing each of their strengths and weaknesses, via "humans are better at/machines are better at" (HABA-MABA) approach as seen in Figure 2.2.

The efforts of separating tasks and allocating them to the one that fits best, human or machine, have been made. Nevertheless, for certain tasks, the performance of humans and machines overlapped. The variable task assignment area is illustrated in Figure 2.3.

As complex tasks, such as medical surgery, require the capabilities of both humans (to make sophisticated judgments) and machines (to perform precise movements), task allocation of to whom the task should be assigned eventually became unclear. Researchers started to drive towards the cooperation of humans and machines to benefit from the distinctive advantages of each side, opening the doors to research in Human-Machine Collaboration. For such collaboration, the system design of taking humans into the machine interaction loop (i.e. human-in-the-loop) is a requisite. The interaction between a human being and a machine is referred to as Human-Machine Interaction which focuses on ameliorating the design and use of interfaces (of computers and agents) allowing human users to interact with them in novel and convenient ways.

2.3.1 Terminology of HAI

HAI replaces the terminology of "machine" with "agent" representing a subdomain of *HMI*. It focuses on the interaction between humans and embodied agents (or Socially Interactive Agents; *SIA*s). In *HMI*, the agent is defined as "anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators" (Russell [2010]). Russell thinks with an Artificial Intelligence (AI) perspective concentrating on the autonomy of entities that compromise various systems. For our work, we take the definition of agents from

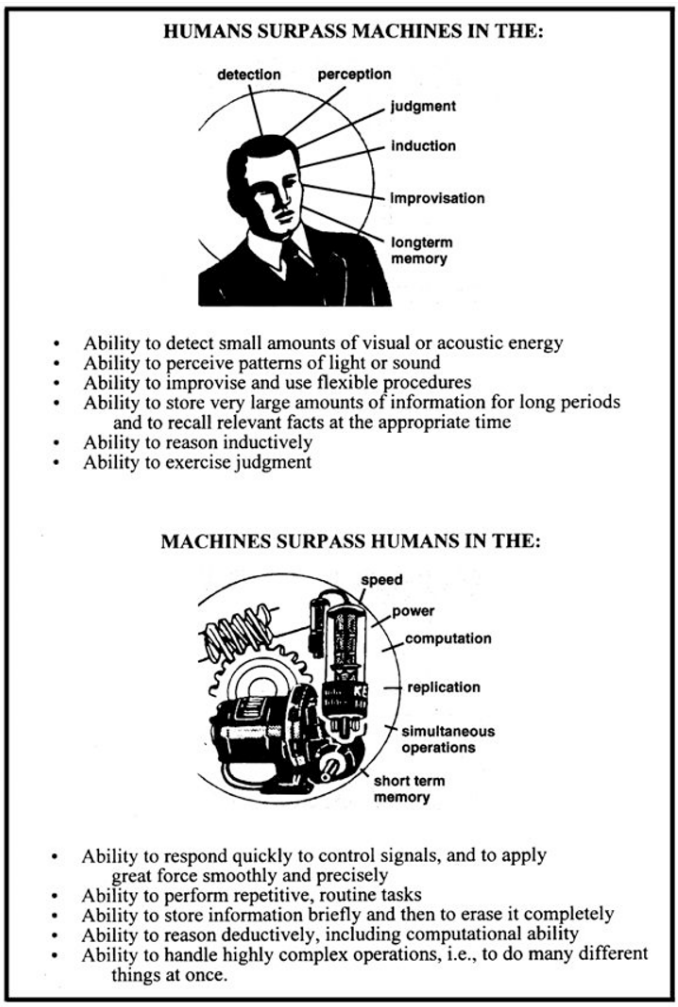


Figure 2.2 HABA-MABA approach of Fitts [1951].

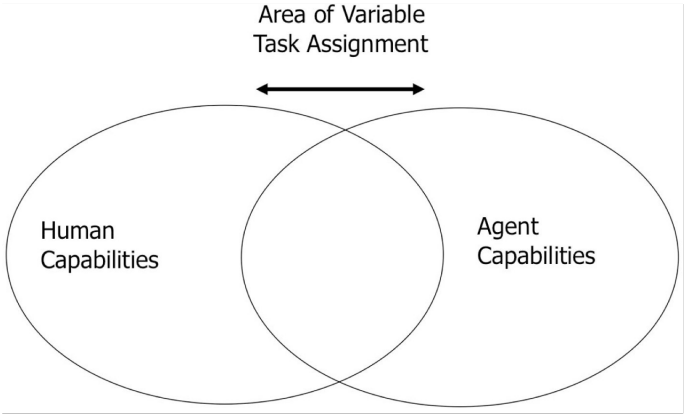


Figure 2.3 Perspective of human and machine/agent capabilities for adaptive allocation and adjustable autonomy (Bradshaw et al. [2017]).

Lugrin et al. [2021] defining them as *SIAs* (or embodied agents) that are capable of autonomously carrying out natural conversations with people in a socially in-

telligent manner. With this definition of agent, we consider *HAI* as the interaction between humans and *SIA*s.

2.3.2 Objective

Like *HMI*, the objective of *HAI* is to improve the interaction between the human and the agent. The advancement of *HAI* focuses on myriad aspects such as engagement (Oertel et al. [2020]), social presence (Pereira et al. [2014], Li [2015]), and behavior realism (Ferstl et al. [2021]) of agents.

Within the human-agent interaction, numerous social signals of diverse modalities are exchanged. As in human-human interaction, the human and agent send and receive multimodal signals which they interpret and use the perceived information for the production of their next behavior. The management of the exchange of multimodal information is also a key aspect of *HAI*.

Chapter 3

Related Work

Contents

In this Chapter, we present an overview of the relevant works addressing our topic of interest of generating *reciprocally adaptive* nonverbal SIA behaviors from *multimodal* signals. We start by introducing the techniques (i.e. computational models) for the different aspects of sequence prediction, nonverbal behavior generation (capturing *intrapersonal* and *interpersonal temporalities*), and multimodal signal processing. As for rendering SIA behaviors, as evaluating them is also important, we describe the evaluation methods employed for *HAI*.

3.1 Sequence prediction techniques

Generating nonverbal behaviors can be considered a similar problem to forecasting future nonlinguistic action sequences. It is thus interesting to investigate existing sequence prediction techniques that could be applicable to nonverbal behaviors.

The methods of sequence prediction can be broadly split into two: offline and online prediction. Offline prediction predicts by giving a sequence of data all at once while online prediction refers to the inference method in which the data are predicted sequentially one after another. We look into these two types to pick the technique that suits best for generating smooth and continuous nonverbal behavior.

3.1.1 Offline prediction

Offline prediction infers with the whole input data given from the start. The entire sequence is predicted at once and this prediction is done in groups (or sequences) and is done independently without considering the previously outputted prediction. Its application can be easily seen for sequence to sequence predictions. Models for such predictions generally have the structure of an autoencoder that encodes the inputted sequence and a decoder that predicts the resulting sequence by decoding the output of the encoder. Sequence to sequence prediction models produce good results for machine translation (Sutskever et al. [2014], Moslem et al. [2023]) and speech recognition (Li et al. [2018], Radford et al. [2023]). The representative models that can be seen in the literature are Bidirectional Long Short-Term Memory (BLSTM; Graves and Schmidhuber [2005]), autoencoders (Greenwood et al. [2017], Ahuja and Morency [2019], Yuan and Kitani [2020]), Generative Adversarial Network (GAN; Goodfellow et al. [2014], Ferstl et al. [2019]), normalizing flow (Henter et al. [2020], Papamakarios et al. [2019]), and Transformers (Vaswani et al. [2017], Bhattacharya et al. [2021], Fares et al. [2022], Radford et al. [2023]).

3.1.2 Online prediction

Unlike offline prediction, online prediction renders the output in a sequential manner predicting for each time-step separately. Among the appliance domains of online prediction, the most representative one is the time series forecasting. Time series forecasting has a wide range of applications such as weather forecasting (Kumar and Jha [2013], Wan et al. [2019]), traffic flow forecasting (Lippi et al. [2013], Tian and Pan [2015]), or stock market prediction (Kim [2003], Tsantekidis et al. [2017]). Various models based on online prediction can be seen in the literature such as Multilayer Perceptron (MLP), Recurrent Neural Network (RNN), Long Short-term Memory (LSTM), Convolutional Neural Network (CNN), and Temporal Convolutional Network (TCN; Palmer et al. [2006], Mohammadi et al. [2018], Tian and Pan [2015], Tsantekidis et al. [2017], Wan et al. [2019]).

Online prediction can be separated into two types which are sliding window prediction and adaptive online prediction. For sliding window prediction, predictions are made independently for each time-step, using only the data of the current

sliding window, and thus without considering the previous output data. Adaptive online prediction also predicts sequentially for every time-step but it uses a memory that is continuously updated during iteration. Thus, the whole sequence of input data is considered to make the prediction at the given time-step and not only the data of a sliding window. This update of the latent vector using the full input data sequence allows adaptive online prediction to render continuous output data.

The output continuity can be further enhanced for both online prediction techniques, sliding window prediction and adaptive online prediction, by integrating the past outputs as input for the future prediction. More precisely, the observations from previous time-steps are fed to a regression equation to predict the value at the next time-step. Such technique that predicts by feeding the output back to the model is called to be autoregressive.

3.2 Nonverbal behavior generation for HAI

The generation of the multimodal behavior of *SIA*s requires modeling the temporality of exchanged social signals. Both *intrapersonal temporality* (coordination of the *multimodal* communicative behaviors within a single person) and *interpersonal temporality* (*multimodal* behaviors arising during dyadic or multi-person interactions) are essential components of the *reciprocal adaptation* as we adapt our behaviors depending on our prior behaviors and the behaviors shown by others.

3.2.1 Intrapersonal temporality

Previous works that modelize *intrapersonal temporality* proposed models that generate facial expressions and communicative gestures linked to speech. They focus on the modeling of *intrapersonal* relationship for a single person (multimodality within the same person). These works employ various Deep Learning (DL) techniques. Several works employ the Feed-Forward Neural Network (FFN; Bebis and Georgiopoulos [1994]) to compute communicative behaviors such as 3D facial animations or head motion from audio (Karras et al. [2017], Ding et al. [2015]). Other models such as the Bi-directional Long Short-Term Memory (BLSTM; Graves and Schmidhuber [2005]) for head movement prediction from sliding temporal windows of prosody features or 3D human body gesture generation from audio utterances (Sadoughi and Busso [2018], Hasegawa et al. [2018]), the conditional autoencoder, the Conditional Variational Autoencoder (CVAE) to predict head pose with speech or to generate diverse body poses via DLow sampling method (Greenwood et al. [2017], Yuan and Kitani [2020]), the Generative Adversarial Network (GAN; Goodfellow et al. [2014]) for generating multiple plausible realizations of real looking communicative gestures or head movements from each speech segment by sampling from a conditioned distribution (Ginosar et al. [2019], Ferstl et al. [2019]), normalizing flow based model (MoGlow method) to speech driven gesticulation generation (Alexanderson et al. [2020]). Recent models employ the Transformers (Vaswani et al. [2017]) to generate emotive body gestures based on text (Bhattacharya et al. [2021]) or face and upper body gestures based on visual,

speech, and linguistic modalities (Fares et al. [2022, 2023]). They have shown that natural and human-like behaviors can be generated by modeling *intrapersonal* relationship.

3.2.2 Interpersonal temporality

Interpersonal temporality is the temporal relationship between participants within an interaction.

The modeling of nonverbal behaviors for dyadic interactions started off with rule-based systems such as manually designed rules that were used for predicting backchannels (Truong et al. [2010]), decision trees for chatbot systems generating natural responses and their timing (Nishimura et al. [2007]), and multimodal probabilistic models that predict backchannels via *multimodal* signals (Morency et al. [2010]). The generation of nonverbal behavior such as facial expression, head and body motion started to flourish with the rise of DL models. Feng et al. [2017] modelized the relationship between a human user and a *SIA*. They generate the agent's facial gestures using the agent's and human's previously predicted facial gestures by creating a Feed-Forward Neural Network (FFN) model. They solely use visual features (facial landmarks) and do not make use of the *multimodal* information present in the interaction. Also, it is exposed to the problem of outputting discontinuous predictions between two time-steps. Grafsgaard et al. [2018] learn by encoding the multimodal signals (facial expression, body motion, and speech) using a Long Short-Term Memory (LSTM) model to predict the facial expression and motion of a partner with the speech of both partners and their facial expression and motion features. The interpersonal relationship is modeled by encoding both partners' behaviors; the multimodality is considered but their behavior predictions risk to be not fluid. Dermouche and Pelachaud [2019b] also study the *interpersonal* relationship by referring it as the interactive loop to generate the agent's behavior. They additionally modelize the *temporality* of nonverbal signals by introducing their Interactive Loop LSTM (IL-LSTM) that considers both agent's and user's upper face behaviors to model the agent's nonverbal behaviors. Similarly, to the model in Feng et al. [2017], the IL-LSTM has the same issue of only taking unimodal input features (facial gestures). As it generates using the sliding window prediction it produces jerky movements and does not consider the whole interaction context. Ahuja et al. [2019] integrates *interpersonal* and *intrapersonal* dynamics via selective attention. They forecast body pose sequences based on the human interlocutor's audio and body pose and audio of the agent.

For motion generation, several works use generative models such as Generative Adversarial Network (GAN; Goodfellow et al. [2014]) and normalizing flow-based models to generate motions that are more diverse and realistic. An extended system of MoGlow (Henter et al. [2020]) is used by Jonell et al. [2020] to predict the agent's facial expression based on the audio of both partners and the facial expression of the human by encoding all modalities using a RNN and passing their concatenation to a neural network at each time-step of the flow. Tuyen and Celiktutan [2022] forecast the upper body motion (face, body, and hand landmarks) with a context-aware model that consists of three components of context encoder, generator, and discriminator. The context encoder encodes the interacting partner's

nonverbal behaviors (body motion and audio) and passes the encoded contextual information to the generator along with the target person's body motion. Then the actions outputted by the generator is injected into the discriminator with the contextual information to validate the motion. The two generative models employed in Jonell et al. [2020], Tuyen and Celiktutan [2022] create various possible behaviors by modeling the two facades of *interpersonal temporality* and *multimodality*. Nevertheless, they face the same problem of not establishing a *continuous link* between two sequentially but separately predicted outputs. With the emerging trend of the Transformers model (Vaswani et al. [2017], Ng et al. [2022]) generate a continuous 3D facial motion of the listener via an autoregressive transformation-based predictor taking the output of the cross-modal attention that combines the speaker's facial motion and audio inputs and that of Vector Quantised Variational AutoEncoder (VQ-VAE; Van Den Oord et al. [2017]) which discretizes the listener's past facial motion. Their architecture allows the modeling of *interpersonal temporality* and *multimodality*, and render *continuous* predictions via the autoregression. One point that could be hindersome about the model is that transformer-like models require massive amount of data to train. Thus, it might not be suitable for all applications that do not have sufficient amount of data.

3.3 Multimodal signal processing

The *multimodality* of signals that can come from words, prosody, and facial expression, is an important aspect that needs to be dealt with for the task of generating nonverbal behavior to ensure an engaging interaction. The works presented in the previous section use *multimodal* signals (audio, visual, and textual features) for nonverbal behavior generation. Nevertheless, they do not all explicitly model *multimodality*. Explicit modeling of multimodal signals can provide a deeper understanding of the exchanged information. Therefore, we observe how these *multimodal* signals can be explicitly processed from models applied for different tasks, including but not limited to nonverbal behavior generation. Chu et al. [2018] propose a neural conversation model generating facial expressions alongside text. Their goal is to add richness to their generation by exploiting modalities in a separate manner. Rather than concatenating both modalities, they use a RNN dedicated to each modality and then obtain the global description by concatenating the history of each modality. Rajagopalan et al. [2016] extended the LSTM for multimodal learning by proposing Multi-View LSTM (MV-LSTM) which explicitly models modality-specific and cross-modality interactions. Thus, the model defines four types of memory cells: modality-specific cells, coupled cells, fully connected cells, and input-oriented cells. MV-LSTM shows promising results (high accuracy for the engagement level prediction task) in exploiting multi-view relationships for behavior recognition. Another approach that learns from multiple modalities was proposed by Zadeh et al. [2018]. Their structure, named Memory Fusion Network (MFN), learns view-specific dynamics in isolation by training a LSTM for each modality and finds cross-view interactions by associating a relevance score to the memory dimensions of each LSTM via an attention mechanism. It stores the cross-view information over time in the Multi-view Gated Memory acting like a

dynamic memory module. MFN has been tested on several multimodal databases and performs highly in sentiment analysis, emotion recognition, and speaker traits recognition. Sharma et al. [2021] modeled an attention-based multimodal for visual question answering of medical images. They use attention modules to focus on the most relevant part of the medical images and questions. Their study shows that the multimodal information can be better captured via the attention mechanism and the interpretability of the results can also be obtained.

Existing models presented above show how we can consider the *temporal coherence* or explicitly model *multimodal* signals. Nevertheless, for the nonverbal behavior generation, these two aspects of *temporality* and *multimodality* are not fully considered. In an interaction both *multimodal* and *temporal* relations of exchanged signals can be observed simultaneously and are correlated. The different modalities provide additional information and the capture of complementary information can be strengthened by explicitly modeling *multimodal* signals. These *multimodal* signals also need to be temporally coherent with each other. The *temporal sync* must be ensured not only between the different modalities of the same person (*intrapersonal temporality*) but also between those of his/her interlocutor (*interpersonal temporality*). Considering the two aspects together can further help capture and understand the correlation between them. It will thus be interesting to investigate how to embed their dynamics for engaging dyadic interactions.

3.4 HAI evaluation

The evaluation of *SIA*'s non-verbal behavior sequences is a difficult and ill-posed problem. We do not display the same behavior all the time. For the same event, depending on various factors such as the person that we are interacting with, the time of day, and our mood, we communicate and react differently to our interlocutor. For example, we may or may not respond to a nonverbal signal (e.g. smile), with more or less intensity and more or less latency. In the same way, head movements are important in maintaining engagement but they do not obey strict and precise laws, and a multitude of movements are possible in response to an interlocutor. However, not all occurring movements are perceived as social, convincing, informative, or even carrying meaningful information. This is what we want to learn during sequence generation: to generate *multimodal* behavior sequences that convey the intended intention (e.g. maintaining engagement) and are perceived as such by the human interlocutors.

3.4.1 Subjective measures

But how do we evaluate the *quality* of the generative behaviors models? There is no unanimous answer to this question today. A large literature on human-agent interaction evaluation has been done to qualitatively measure behavior quality. Various aspects of the generated behaviors are assessed which are:

- Behavior naturalness (Fitrianie et al. [2021], Von der Pütten et al. [2010]): e.g. "Is the behavior of the virtual agent artificial?";

- Behavior human-likeliness (Fitrianie et al. [2021], Von der Pütten et al. [2010]): e.g. "Does the virtual agent behave like a human?";
- Engagement (Fitrianie et al. [2021], Von der Pütten et al. [2010]): e.g. "The virtual agent was engaged in the conversation?";
- Synchrony (Prepin et al. [2013], Louwerse et al. [2012]): e.g. “The virtual agent and I were agreeing to each other?”;
- Rapport (Wang and Gratch [2009], Von der Pütten et al. [2010]): e.g. "I think the virtual agent and I established a rapport.".

3.4.2 Objective measures

While some studies focus on evaluating only based on subjective study (Jonell et al. [2020]), objective measures are also vital to fully assess the quality. We present here some quantitative measures used in the literature (Feng et al. [2017], Dermouche and Pelachaud [2019b], Grafsgaard et al. [2018], Ng et al. [2022]).

Behavior precision

Like any other regression model evaluation, we could accept a generated behavior to be correct based on its quantitative closeness to the ground truth. In other words, the accuracy is calculated using metrics such as Mean Squared Error (MSE; Ding et al. [2015], Sadoughi and Busso [2018]), Root Mean Squared Error (RMSE or L2; Feng et al. [2017], Dermouche and Pelachaud [2019b], Ng et al. [2022]), Mean Average Error (MAE; Ginosar et al. [2019]), and Average Position Error (APE; Hasegawa et al. [2018], Ahuja and Morency [2019], Ahuja et al. [2019]) for each sample.

Behavior likelihood

Various outcome behaviors could derive from the same surrounding signals. There is not a fixed behavior (the ground truth) for a given situation and there are multiple plausible answers. A lot of solutions are used to estimate the quality of sequences generated using Generative Adversarial Network (GAN). They are often based on the principle that several sequences are generated for the same testing example. A solution consists of estimating the distribution over generated sequences and then, calculating the log-likelihood of the ground truth sequence (Sadoughi and Busso [2018], Jonell et al. [2020], Mao et al. [2021]). Another solution is to measure the smallest distance between the generated sequences and the ground truth one, and average these distances along the testing sequences. Aliakbarian et al. [2021] estimate the diversity of the generated sequences as the average distance between all pairs of generated sequences. At the same time, they measure the quality using a binary classifier that discriminates between real and generated sequences. Other authors use statistical measures of Inception Score (ID) or Frechet Inception Distance (FID) to measure the generation fidelity of human motion (Aliakbarian et al. [2020], Cai et al. [2021]).

Behavior Synchrony

All the previous measures assume that several sequences are generated for the same test sequence or that we can estimate the distribution of real sequences. The reciprocal adaptation leads us to a very specific case where the previous measures cannot be applied. The mutuality of the adaptation must be evaluated and the whole interaction must be taken into account. More importantly, a lot of temporal dependencies exist between both partners and these phenomena are not observed using the previous measures. Thus, we are also interested in the interpersonal relationship and how to measure it.

While conversing, the speech and movement of the interlocutors are dynamically coordinated (i.e. interpersonal synchrony). However, the detection of such coordination is not so simple. In a real conversation, the signals do not always happen simultaneously at the same moment. The signals generally are exchanged one after another. The most common way to measure interpersonal synchrony is via Pearson's correlation (PCC; Campbell [2008], Delaherche and Chetouani [2010], Reidsma et al. [2010], Zadeh et al. [2018], Ng et al. [2022]).

Each interlocutor can send or respond to a signal with a certain time delay (after a perception time (Chartrand and Bargh [1999])). For example, when a person smiles, the interacting person can respond to this smile or not. This response is perceived as a mimic of the first smile if it happens within a time delay of 2 to 4 seconds (Leander et al. [2012]). Thus, we need to take into account time shifts. Several works address this by applying the time-lagged cross-correlation (TLCC; Boker et al. [2002], Ashenfelter et al. [2009], Beňuš et al. [2011]). A hindersome limitation of correlation is that a window length of interaction must be chosen to perform the correlation. However, the window sizes can vary for each produced motion and are not the same for interactors.

Another method of synchrony evaluation is the recurrence analysis (Shockley et al. [2003], Varni et al. [2010]). The analysis assesses "recurrence points" which are points in time where similar states (or patterns of change) are visited by two different systems. The recurrent analysis depends on manipulable states (e.g. posture state or affect state) and shows a graphical representation (a diagonal structure) of time periods when two systems visit the same state. For the recurrent analysis, the evaluation requires a fixed length of system periods and time shifts.

The mimicking behaviors can also differ in terms of duration and intensity. This implies that the sequence comparison also needs to be invariant to dilations when comparing the signals. A well-known technique that deals with such aspects is Dynamic Time Warping (DTW; Müller [2007]). The similarity between two temporal sequences of different speed and length can be measured.

New indicators characterizing synchrony phenomena were introduced by (Rauzy et al. [2022]). They consider the two signal timescales as oscillating normal modes associated with the sum and the difference of the trajectories (x_{sum} for symmetric mode and x_{diff} for asymmetric mode). Based on the two, they propose new indicators (mode characteristic periods, coupling factor, coefficient of synchrony, and energy) to evaluate the synchrony.

Model	Interpersonal features	Modalities	Continuity	Small dataset
Karras et al. [2017]		a, e		✓
Alexanderson et al. [2020]		a, v, s	✓	✓
Fares et al. [2023]		a, v, t		
Feng et al. [2017]	✓	v		
Grafsgaard et al. [2018]	✓	a, v		
Dermouche and Pelachaud [2019b]	✓	v		✓
Jonell et al. [2020]	✓	a, v		✓
Tuyen and Celiktutan [2022]	✓	a, v		
Ng et al. [2022]	✓	a, v	✓	
Our proposition	✓	a, v	✓	✓

Table 3.1 SIA behavior generation models. Input modalities marked as a:audio, v:visual, t:text, e:emotion, and s:style.

As an alternative to temporal methods, spectral analysis was suggested. The evolution of relative phase for a stable time-lag between interlocutors is measured (Oullier et al. [2008], Richardson et al. [2007]). It also renders information about the coordination stability with the flatness degree of the phase distribution and the overlapping frequency via the cross-spectral coherence. The synchrony can be also measured in the time-frequency domain via cross-wavelet coherence (Hale et al. [2020]).

3.5 Discussion

The generation of nonverbal signals is *time-dependent* like time series problems. The memory retention present within recurrent networks such as RNN, LSTM, and TCN, has shown great promise in time series forecasting. As human behaviors heavily depend on previously performed ones, this aspect of memory is also important for our situation. Moreover, as behavior must be *continuous*, it is preferable to employ adaptive online prediction along with the aspect of predicting based on the previous time-stamped data in an autoregressive manner (ref. Section 3.1).

The aforementioned models (ref. Sections 3.2 and 3.3), resumed in Table 3.1, show how the relationship between the social signals of oneself (*intrapersonal* relationship), the signals of the interlocutors (*interpersonal* relationship), and the *multimodal* signals can be modeled. For our work, we want to model the *reciprocal adaptation* by considering the two facets of *temporality* (both *intrapersonal* and *interpersonal*) and *multimodality* along with the *continuity* aspect for the generation of our agent's nonverbal behavior. The *multimodality* modeling is absent in Feng et al. [2017], Dermouche and Pelachaud [2019b] and the *continuity* is not assured for Feng et al. [2017], Grafsgaard et al. [2018], Dermouche and Pelachaud [2019b], Jonell et al. [2020], Tuyen and Celiktutan [2022]. While Ng et al. [2022] meets all three of our criteria, it requires a lot of training data. In our case, we use a small database (ref. Chapter 4), making their model not suitable for our application. We propose a new model, Augmented Self-Attention Pruning (*ASAP*)

model presented in Chapter 6, that renders *continuous* nonverbal behaviors (for both *speaker* and *listener*) performing with a small dataset. It also learns to capture the *interpersonal* relationship between the interlocutors from the exchanged *multimodal* signals to endow *SIA*s with the *reciprocal adaptation* capability. Moreover, we develop another model, Historical Intrapersonal Interpersonal ADAptive Multimodal model (*HI*^{2}*-ADAM*) model detailed in Chapter 8, which also captures the *reciprocal adaptation* to generate *adaptive* and *continuous* nonverbal *SIA* behavior (for both roles) like the *ASAP* model. *HI*^{2}*-ADAM* model better encodes the adaptation between the interlocutors by explicitly modeling the *intrapersonal* relationship with the modality histories (*modality memory*) and a deeper encoding of the *multimodal* signals.

Various efforts have been made to quantify the quality of nonverbal behaviors (ref. Section 3.4). Nevertheless, there is not yet a perfect metric to evaluate them. Especially several aspects of behavior quality such as naturalness and human-likeness might be trivial for a human, but still very hard to access for a machine (Fitrianie et al. [2020, 2021]). Thus, human evaluation remains a critical part of behavior evaluation (Feng et al. [2017], Karras et al. [2017], Chu et al. [2018], Sadoughi and Busso [2018], Alexanderson et al. [2020], Jonell et al. [2020], Yuan and Kitani [2020], Cai et al. [2021], Fitrianie et al. [2020]). To better evaluate the *interpersonal synchrony* between the human and the agent and to complement the subjective evaluation, we propose the use of new metrics for agent behavior evaluation, presented in Chapter 6, and new *reciprocal adaptation measures* (synchrony and entrainment loop measures), introduced in Chapter 5.

Chapter 4

Corpus

Contents

This Chapter presents the corpus that we chose to use for our study of generating reciprocally adaptive behavior for human-agent interaction (consisting of analysis in Chapter 5, reciprocal adaptation modeling in Chapters 6 and 8, and real-time system in Chapter 7). We also explain the applied processes of feature extraction and data processing along with the employed annotations.

4.1 NoXi Corpus

For our study, we use the NoXi database (Cafaro et al. [2017]), shown in Figure 4.1, which is a corpus of screen-mediated face-to-face interactions. It contains natural dyadic conversations talking about a common topic. Each interacting dyad consists of a pair of participants with two different roles which are called expert and novice. The expert is the one who transfers information with the goal of sharing his/her knowledge on a topic and thus who leads the conversation by talking more frequently and for a longer time. The novice (the other interacting partner) receives the information and responds to the sayings of the expert on the topic.

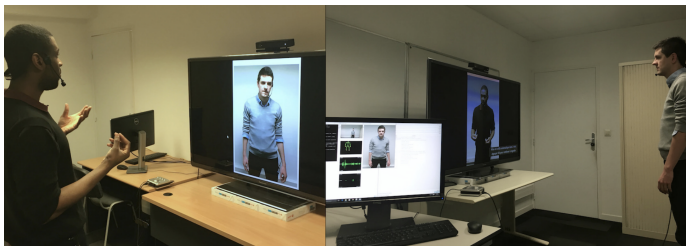


Figure 4.1 Snapshots of the NoXi corpus's recording session.

The NoXi corpus consists of 3 parts depending on the recording location (France, Germany, and UK). For our work, we only use the recording from the French location which consists of 21 dyadic interactions performed by 28 participants (23 males and 5 females) with a total duration of $7h22min$. The participants' age is in the range of $18 - 45$ years old mainly consisting of $21 - 25$ years.

4.2 Feature Extraction

We obtain nonverbal behavior features for both interacting participants through feature extraction. For each time-step, the visual and audio features, listed in Table 4.2, are extracted using opensource toolkits of OpenFace (Baltrušaitis et al. [2016]) and openSMILE (Eyben et al. [2010]) (after a denoising phase explained below) respectively.

4.3 Data processing

To clean up the data, we process the visual and audio features separately.

4.3.1 Visual data processing

We process the visual data on the extracted visual feature by firstly filtering out unsuccessful extractions (with a success rate under 0.7) as there are some moments where OpenFace extracts the features with low accuracy extracting wrong feature values. We interpolate (linear interpolation) to fill in the dropped values

Feature			Description
Visual	Head Rotation	R_x	Head rotation around the x axis
		R_y	Head rotation around the y axis
		R_z	Head rotation around the z axis
	Head Translation	T_x	Head translation around the x axis
		T_y	Head translation around the y axis
		T_z	Head translation around the z axis
	Gaze (or Eye Movement)	G_x	Gaze around the x axis
		G_y	Gaze around the y axis
	Facial Action Units (Facial AUs detailed in Appendix A in Chapter 10; Ekman and Friesen [1976])	$AU1$	Inner brow raiser
		$AU2$	Outer brow raiser
		$AU4$	Brow lowerer
		$AU5$	Upper lid raiser
		$AU6$	Cheek raiser
		$AU7$	Lid tighten
		$AU12$	Lip corner puller for smile
		$AU15$	Lip corner depressor
Audio	Fundamental Frequency (or Pitch)	$F0$	Predominant frequency representing the speech quality
	Loudness		Speech intensity from the auditory spectra
	Voicing probability		Speech presence probability expressed as a probability score in the range of 0 to 1
	Mel-Frequency-Cepstral Coefficients (MFCC; Logan [2000])		Representation of the short-term power spectrum of a sound; represented by 13 $MFCC$ features (0-12)

Table 4.2 Extracted features from the NoXi corpus.

and smooth out the feature data by applying a median filter (with a window size of 7 found after manual verification). Then, different scaling techniques, normalization (shifting and rescaling values to make the data range between 0 and 1; or Min-Max scaling) and standardization (setting the attribute mean to 0 and the distribution to have a unit standard deviation) methods depicted in Table 4.3, are applied depending on the characteristic of each feature. The scaling technique of Type 1 (a standardization method) is applied to the head rotations ($R_{x,y,z}$), head translations ($T_{x,y,z}$), and gaze ($G_{x,y}$) as their values are centered around the mean with a unit standard deviation. For facial AUs (*AU1*, *AU2*, *AU4*, *AU5*, *AU6*, *AU7*, *AU12*, and *AU15*), the Type 2 scaling technique (a normalization method) is employed as their intensity values ranges from 0 to 5.

4.3.2 Audio data processing

The audio processing starts with the denoising of the audio files, removing surrounding sound, via Audacity's noise reduction (Audacity [2017]). The denoised audio is used to extract the audio features. The extracted audio features are processed with their corresponding scaling technique similar to the preprocessing of visual features as found in Table 4.3. The applied scaling techniques for different audio features are as follows. Type 1 is applied to the 12 *MFCC*s (1-12) showing data centers around a mean and different standard deviations of which we want to scale to 1. Type 2 is applied to the pitch ($F0$) ranging from 0 to each speaker's maximum pitch level. Type 3 (a normalization method) is applied to loudness also ranging from 0 to each speaker's maximum loudness value with different standard deviations. To respond to the difference in standard deviation, the standardization method (Type 3) includes a multiplication with α (feature standard deviation coefficient) which allows the different values of the same feature to share the same standard deviation value. Type 4 (a normalization method) is applied to *MFCC*0 consisting of values varying between a minimum and maximum value.

All the features are adjusted to $25fps$ for our study.

	Scaling Technique	Feature(s)
Type 1	$\frac{f-\mu}{\sigma}$	$R_{x,y,z}$, $T_{x,y,z}$, $G_{x,y}$, and $MFCCs$ (1-12)
Type 2	$\frac{f}{f_{max}}$	$AUs(1,2,4,5,6,7,12,15)$ and $F0$
Type 3	$\frac{f}{f_{max}} * \alpha$	Loudness
Type 4	$\frac{f-f_{min}}{f_{max}-f_{min}}$	$MFCC0$

Table 4.3 Scaling technique types corresponding to each feature. f corresponds to the feature values, f_{max} corresponds to the maximum feature value, f_{min} corresponds to the minimum feature value, μ corresponds to the mean of the feature values, σ corresponds to the standard deviation of feature values, and α corresponds to the feature standard deviation coefficient.

4.4 Annotations

To analyze the human-agent interaction, we not only look into low-level signals (i.e. extracted features) exchanged within human-human interactions of the NoXi corpus but also study the high-level signals that are annotated. Annotations of engagement and social dimensions of warmth and competence, annotated at the signal level, are available for the NoXi corpus (available with the annotation tool NOVA (Heimerl et al. [2019]) shown in Figure 4.2).

For the engagement annotations, the perception change of engagement was characterized in Dermouche and Pelachaud [2019a] with five levels (0: strongly disengaged; 1: partially disengaged; 2: neutral; 3: partially engaged; 4: strongly engaged). In Biancardi et al. [2017], the continuous annotations of social dimensions of warmth and competence were done with scores ranging from 0 to 1 (0: very low degree of perceived warmth or competence; 1: very high degree of warmth or competence). As the work of Biancardi et al. [2017] focuses on the

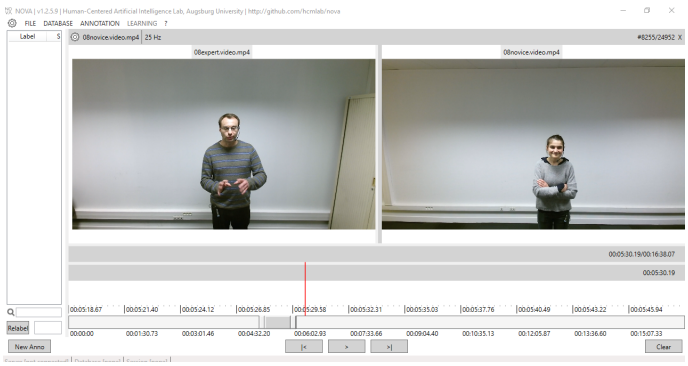


Figure 4.2 NOVA annotation tool.

expert, we also use the annotations on the three aspects of engagement, warmth, and competence of the expert.

We also get the annotations of the conversational state automatically by performing voice activity detection (VAD), which is a binary classifier that detects the presence of human speech in audio, on the denoised audio files.

Part II

Human-Agent Interaction Analysis

Chapter 5

Human-Agent Interaction Analysis

Contents

In this Chapter, we present the analysis of a corpus of dyadic human-human interactions around adaptation. With the goal to better analyze human-agent interactions, we propose new *reciprocal adaptation measures*.

5.1 Introduction

Research around *SIA*s mostly focuses on the development of creating a more *social*, *engaging*, and *human-like SIA*s. Not only is it important to develop such *SIA*s but it is also essential to evaluate their quality. Nevertheless, the task of evaluating the agent's quality is challenging. In particular, the way of objectively evaluating human-agent interactions is not evident. As *SIA*s are interacting with the human users, the assessment must not only be done at the agent's side but also at the interaction level considering the human interlocutor. To address this problem, we propose new *measures of adaptation* by looking into real human-human interactions. Among the different aspects of adaptation, we focus on assessing the synchrony and entrainment loop of dyads.

As a first step in studying our adaptation measures, we observe how synchrony and entrainment loop participate in the perception perspectives of engagement between interlocutors and interlocutors' social attitudes. We hypothesize to see a relationship between *reciprocal adaptation* (synchrony and entrainment loop) and engagement levels. We also hypothesize that *reciprocal adaptation* may have an impact on the perception of the social dimensions of warmth and competence of the interlocutors.

In our study, we focus on smile, a social signal that may convey a great variety of communicative and emotional functions (Niedenthal et al. [2010], Hess et al. [2014]). Smiles are frequently observed during an interaction (Knapp et al. [2013]). They can signal friendliness and positive emotions; they can be used as a polite signal to greet an acquaintance; they can be indicated as agreement and liking; etc. Smile is an important socio-emotional signal that has received a lot of interest in affective computing domains. Previous studies have highlighted the power of smiling *SIA*s to achieve such a goal (Wang and Ruiz [2021], Ochs and Pelachaud [2013]).

We present new *reciprocal adaptation measures* that can be employed to objectively evaluate the quality of the agent in human-agent interaction. Our ultimate goal is to build socially interactive SIAs that are able to maintain user engagement during an interaction. In the scope of this section, we are interested in studying the *reciprocal adaptation* of smile behaviors in dyadic interactions. To do so, we propose new objective measures that study the synchrony of behaviors including their absence of response and behavior entrainment loop to better understand how nonverbal behavior adaptation emerges during an interaction. We aim to investigate how they are displayed between the participants of an interaction and how they participate in the perception of conversational engagement and social attitudes of the participants. We look at the relation of *reciprocal adaptation* with the engagement level and the social dimensions of warmth and competence.

5.2 Related Works and Limitations

During a conversation, interlocutors dynamically adapt by coordinating their speech and behaviors (Condon and Ogston [1967], Burgoon et al. [1995], Bernieri and Rosenthal [1991], Chartrand and Lakin [2013]). Among the various

social signals that are produced during an interaction, the smile is one of the most important human interaction signals. The smile alone can express diverse information (e.g. affect state, level of engagement, and intrinsic nature) to the interacting partner in a variety of social contexts (Ekman [1992], Hess et al. [2002]). The presence of a smile that incorporates such diverse implications can impact the perception by other partners (e.g trust, intelligence, warmth, and attractiveness) (Scharlemann et al. [2001], Lau [1982], Reis et al. [1990]). As such, we want to check the influence of smiles between the interlocutors and thus hold an interest in measuring the smile adaptation. To find out how to measure the adaptation of smiles, we investigate related measures notably synchrony measures (e.g. measures for nonverbal signals and biomedical signals).

Early works on synchrony started off with manual assessment done by trained observers who were trained to perceive it directly in the data. Such evaluations were based on behavior coding methods that evaluate the interaction behaviors on a local scale by analyzing them in micro-units (Cappella [1997], Condon and Sander [1974]). However, the training of observers is very labor-intensive which led them to switch to a judgment method that uses a Likert scale to rate behaviors on a longer time scale (Cappella [1997], Bernieri et al. [1988]). The problem with manual annotations, that rely on perception by a third party, is that it is very costly. Manual annotations are very time-consuming and there is a risk of being biased as the label decision depends heavily on the annotator. Thus, we want an objective evaluation technique that can automatically process and render a non-biased synchrony measure.

Automatic measures enable us to avoid the tedious work of manual annotation by automatically capturing relevant social signals that detect the presence of synchrony. Measures that are frequently employed for interpersonal synchrony are Pearson's correlation (PCC; Campbell [2008], Delaherche and Chetouani [2010], Reidsma et al. [2010], Zadeh et al. [2018], Ng et al. [2022]), time-lagged cross-correlation (TLCC; Boker et al. [2002], Ashenfelter et al. [2009], Beňuš et al. [2011]), and recurrence analysis (Shockley et al. [2003], Varni et al. [2010]). Nevertheless, to perform such measures, a fixed window size is necessary. This may be problematic as produced motions may vary in length and do not happen exactly after a certain time but within a time delay (e.g. 2 to 4 seconds) (Chartrand and Bargh [1999], Leander et al. [2012]).

The response of a smile is very dynamic. Each smile is not produced with the same length, and as stated above, the timing of the smile varies. For example, when we are asked to reproduce a smile that we have made, it is almost impossible to recreate the exact same smile with the same duration and timing. To address such dynamics, the measure must be invariant to dilations and shifts. A frequently used technique to do so is the Dynamic Time Warping (DTW) (Müller [2007]) which assesses the similarity between two temporal sequences of different speed and length. Nevertheless, the DTW matches every index of a sequence with one or more indexes from the other, which can be problematic for our case of nonverbal behaviors as both cases of a behavior occurring or not are correct answers (i.e. absence of response, for instance a person can reply with a smile or choose to not reply but both cases are plausible responses) but the DTW will consider it as an error.

The field of biomedical signal processing also holds a big interest in such synchrony measures for applications such as detecting synchrony in EEG (Bakhshayesh et al. [2019]). Various metrics are employed from point to point measures such as correlation and coherence (a linear correlation computed in the frequency domain via cross spectrum), correntropy coefficient (a correlation measure that is sensitive to nonlinear relationship and high order statistics), wav-entropy coefficient (a correntropy computed in the time-frequency domain with wavelet transforms), to measures that are solely focused on synchronization like phase synchrony (an amplitude-independent estimation of signal phase relationship) and event synchronization (a measure calculated from the number of occurrences of predefined signal events, counting events that are followed by another event in the other signal within a specified time, and their symmetric counterpart). Yet these measures are not suitable for our use as stated above for point to point measures because the subsequences of a signal might have different phase delays which could be troublesome.

In our work, we are interested in measuring how people adapt their behavior, in particular their smiles, during an interaction. During an interaction, participants may respond and adapt to each other's behavior. These interactive behaviors may serve not only to reinforce the relationship between the participants (their engagement in the interaction) but also to display different social attitudes. We are interested in measuring *reciprocal adaptation* as a function of synchrony patterns and entrainment loop. Our measure of synchrony patterns includes when participants respond or not to each other's behaviors. The absence of response is considered as an error by the point to point measures (e.g. correlation) and the DTW approach and is completely ignored by the recurrent analysis, spectral analysis, and cross-wavelet analysis. However, the absence of response may also convey important information about the interaction. In order to study the impact of the absence of response, we need to introduce a new measure.

5.3 Synchrony measures

To our knowledge, existing measures, mentioned above, are not suitable for our problem, notably regarding the absence of a response and capturing the entrainment loop. To overcome this limitation, we propose a new way to measure the reciprocal adaptation for a dyadic pair that measures the synchrony of behaviors including their absence of response while tolerating time swift, dilation, deletion, and insertion, and capture the behavior entrainment loop. This measure is also able to detect the addition (produced by oneself without the reaction of the other) and the suppression (produced by the other without the reaction of oneself) of signals as illustrated in Figure 5.1.

5.3.1 Definition

We first address the problem by taking into account the absence of response when measuring synchrony. Our method derives from the classical sequence dissimilarity quantification technique called edit distance or Levenshtein distance (Navarro

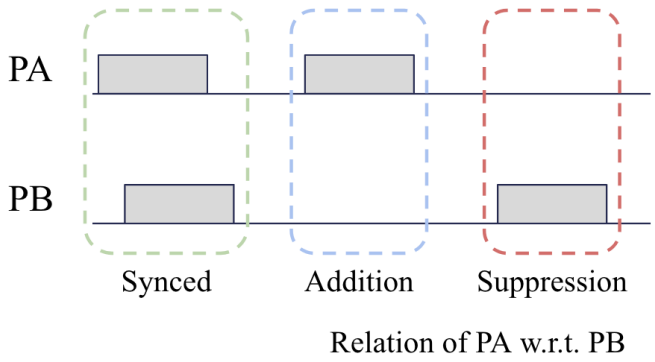


Figure 5.1 Illustration of synced pairs and unsynced pairs (i.e. addition and suppression).

[2001]). Its use can be mostly observed in fields such as natural language processing (Lhoussain et al. [2015]) and bioinformatics (Chang and Lawler [1994]) as it compares the similarity between two strings (e.g. words) by counting the minimum number of transformation operations that are required to convert one string into the other. We grab the concepts of insertion and deletion of the edit distance while we don't use the concept of substitution.

We evaluate the synchrony with signal activation by converting continuous values to binary values and extracting subsequences corresponding to active signal parts, with their starting (s) and ending (e) times. We choose to binarize the continuous values to better see the impact of absence of response. Let us consider an active subsequence (sequence of 1) A from person PA and B from person PB.

To detail, we use the term smile to refer to $AU12$; though we are aware thaa t smile may be produced by different facial AUs (e.g. $AU11$, $AU13$...) in combination with other AUs (such as $AU6$ or $AU1$, $AU2$; Ekman and Friesen [1982]). The analysis studies the smile activation values which are obtained by binarizing the continuous intensity value of smile ($AU12$) with the threshold of $1.5/5$ which is the minimal intensity (manually identified) for a smile activation.

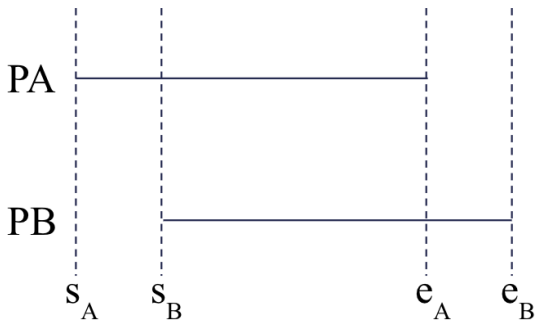


Figure 5.2 Illustration of two subsequences.

We consider that both subsequences are synchronized or paired (i.e. synced pair) if:

$$|e_A - e_B| + |s_A - s_B| \leq threshold \tag{5.1}$$

where s_A and s_B are the starting time points and e_A and e_B are the ending time points of subsequences A and B respectively, and the threshold is set to twice the mimicry time delay. The subsequences are illustrated in Figure 5.2. For our application of measuring the synchrony of smile, we took the threshold of 4 seconds (considering all responses that happen within a maximum of 4 seconds); actually the literature on nonverbal behavior mimicry states that the mimicry time delay can vary from 2 to 4 seconds (Chartrand and Bargh [1999], Leander et al. [2012]).

If several subsequences of a person check this condition with the same subsequence of the other person, a synced pair is formed with the one that has the minimum distance. The other subsequences are not paired.

Both paired subsequences and unpaired subsequences of persons A and B, illustrated in Figure 5.1, are considered to estimate the synchrony:

$$PA\&PB = \frac{nb.\ of\ synced\ pairs}{total\ nb.\ of\ events}$$
$$PA\&\neg PB = \frac{nb.\ of\ unpaired\ subseq.s(seqA|seqB)}{total\ nb.\ of\ events}$$
$$PB\&\neg PA = \frac{nb.\ of\ unpaired\ subseq.s(seqB|seqA)}{total\ nb.\ of\ events}$$

where the total number of events is the sum of the number of synced pairs and the number of unpaired subsequences (i.e. unsynced pairs corresponding to addition and suppression cases) of both persons A and B.

Each measure renders a probability that corresponds to:

- $PA\&PB$: PA and PB responding to each other,
- $PA\&\neg PB$: PA is active but not PB,
- $PB\&\neg PA$: PB is active but not PA.

$PA\&PB$ means that both participants smile simultaneously or with a small delay corresponding to the reacting time; this measure represents the sync between PA and PB. For $PA\&\neg PB$ and $PB\&\neg PA$, only one of the person is acting (PA smiles and PB does not smile, and vice versa), these measures indicate that PA and PB are not in sync.

5.3.2 Analysis

Smile Statistics To start off, we wanted to visualize the statistics of smiles in terms of their occurrence frequency and duration in our database depending on the person's role (expert or novice). We annotate Person 1 (novice in the NoXi database; ref. Chapter 4) as $P1$ and Person 2 (expert in the NoXi database; ref. Chapter 4) as $P2$.

With the visualization of the smile occurrence statistics in Figure 5.3 (left), we note that $P1$ tends to smile more often than $P2$. The context of the dyadic interaction of the NoXi corpus is mainly friendly and positive. Participants were paired between one who wanted to talk about a topic and one who wanted to learn about this topic (Cafaro et al. [2017]). Within such an interaction context, having $P1$

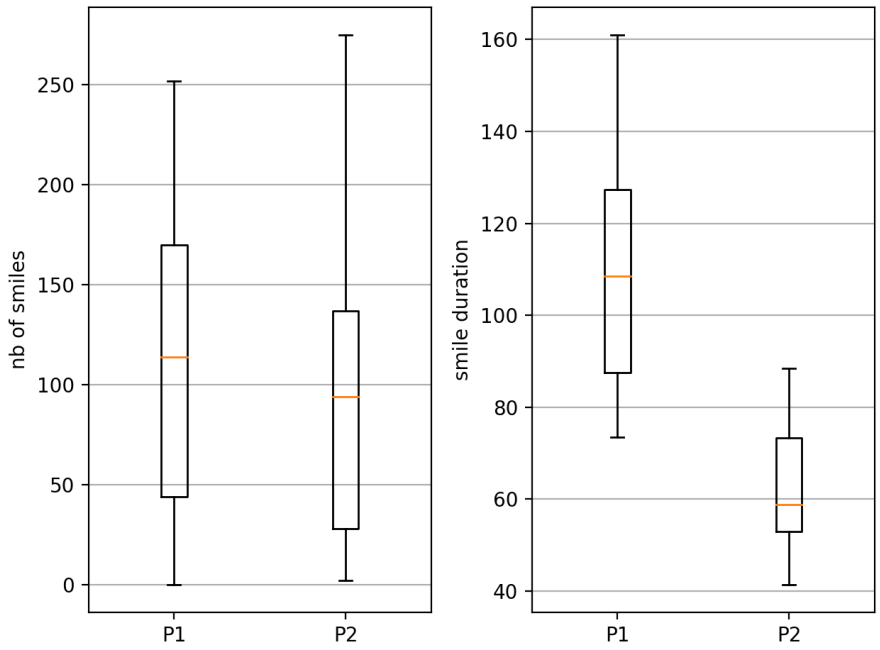


Figure 5.3 Number of smiles produced by $P1$ and by $P2$ (left) and Smile durations of $P1$ and $P2$ (right).

smiling more than $P2$ can be explained by $P1$ displaying positive backchannels or showing actively his/her involvement when $P2$ is talking. Along with the number of smiles produced by the participants, we also hold interest in the smile duration statistics. Figure 5.3 (right) shows that $P1$ generally maintains his/her smile longer than $P2$. This can further support our analysis that $P1$'s smiles may have the purpose of showing conversational involvement.

Smile Synchrony Statistics Going back to our initial objective of investigating the reciprocal adaptation of smile and its relation with the perception of social attitudes, we start by analyzing the smile with our measures of synchrony behaviors including their absence of response.

We computed the probability densities, via our proposed measures, to visualize the distribution of 3 cases: $P1$ and $P2$ responding to each other ($P1\&P2$), $P2$ smiling to $P1$ but not reversely ($P2\&\neg P1$), and $P1$ smiling to $P2$ but not reversely ($P1\&\neg P2$).

We can remark, in Figure 5.4, that during the conversation both $P1$ and $P2$ produce smiles that are in sync responding to one another (smiling at the same time or following back within the mimicry delay of 4 seconds) and also smiles that are not responded by the other partner. As seen in Figure 5.3, $P1$ has a higher probability of smiling even during the absence of the other interacting partner's response ($P1\&\neg P2$), because of his/her tendency to smile more than $P2$.

Synchrony Clustering To better investigate the synchrony between the two interlocutors, we decided to first check if the smile synchrony of the 21 video dyads of

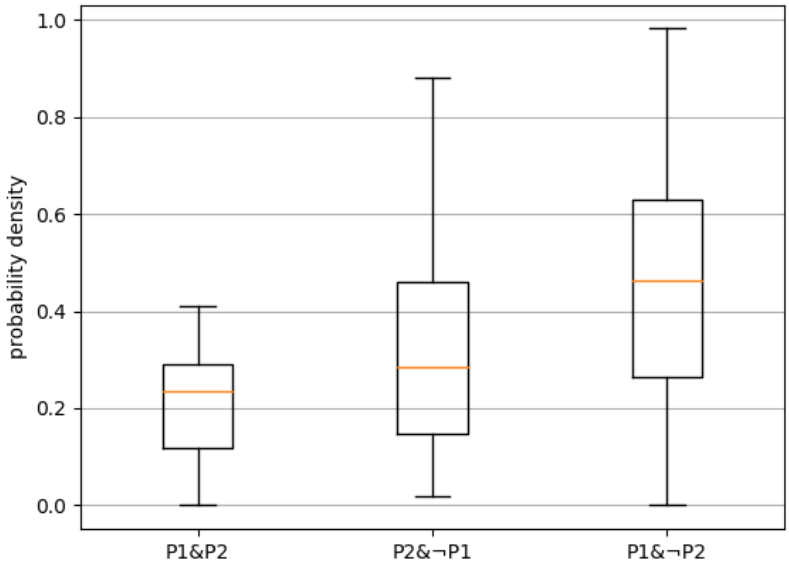


Figure 5.4 Probability of smiles that are in sync ($P1\&P2$), $P2$ smiling without the response of $P1$ ($P2\&\neg P1$) and $P1$ smiling without the response of $P2$ ($P1\&\neg P2$).

the NoXi corpus can be classified into different groups. We performed a dendrogram hierarchical clustering to cluster the dyads using our obtained measures of synchrony behaviors including their absence of response ($P1\&P2$, $P2\&\neg P1$, and $P1\&\neg P2$). As seen in Figure 5.5, we split our data into three clusters by cutting the dendrogram with a threshold of 1.0. The cluster classes can be visualized in the 3-dimensional space of our proposed measures in Figure 5.6.

In Figure 5.7, we can note in cluster 1 that a low synchronization level ($P1\&P2 \sim 0.072$) occurs along with when $P2$ smiles very frequently ($P2\&\neg P1 \sim 0.924$) while $P1$ does not smile much ($P1\&\neg P2 \sim 0.004$). A medium synchrony ($P1\&P2 \sim 0.231$) is seen, in cluster 2, when $P1$ smiles a lot ($P1\&\neg P2 \sim 0.637$) and $P2$ smiles a bit ($P2\&\neg P1 \sim 0.146$). For cluster 3, a high synchrony ($P1\&P2 \sim 0.33$) is observed when $P1$ and $P2$ both smile frequently ($P2\&\neg P1 \sim 0.408$ and $P1\&\neg P2 \sim 0.305$).

We can deduce from these three clusters that the highest level of synchronization ($P1\&P2 \sim 0.33$) is correlated with both interacting partners who tend to smile frequently, while the lower levels of synchronization, in cluster 1 ($P1\&P2 \sim 0.072$) and cluster 2 ($P1\&P2 \sim 0.231$), are correlated with the situation when one of the partners, independent of his/her role, does not respond much. This shows how the presence of smile reciprocity is an important factor with respect to synchrony level ($P1\&P2$); a partner that nearly does not respond to other's smile ($P1\&\neg P2 \sim 0.023$ in cluster 1) deteriorates the synchrony of the two even when the other interlocutor smiles a lot ($P2\&\neg P1 \sim 0.924$ in cluster 1). It confirms that synchronization is highly dependent on coordination between partners (Burgoon et al. [1995], Tschacher et al. [2014]).

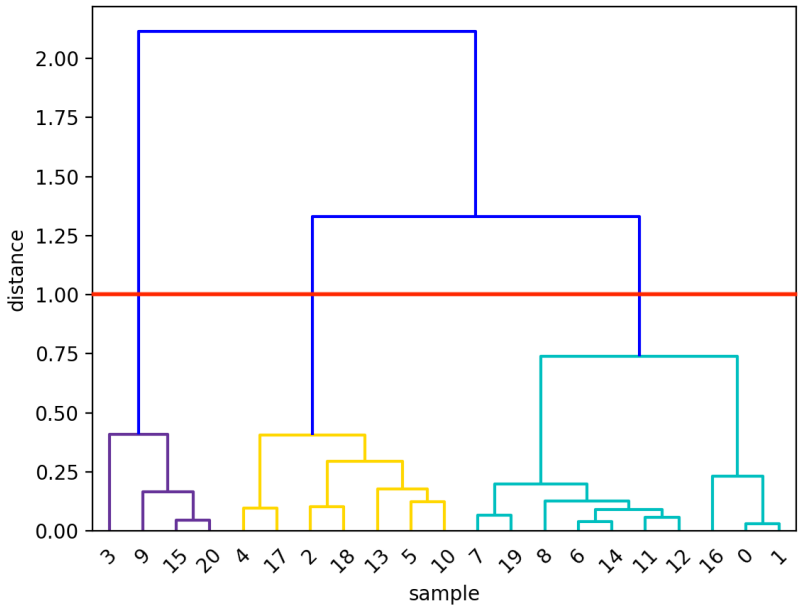


Figure 5.5 Dendrogram of synchrony measures where the distance is the distance between the sample points in the 3D space of our proposed measures of synchrony. Threshold of 1.0.

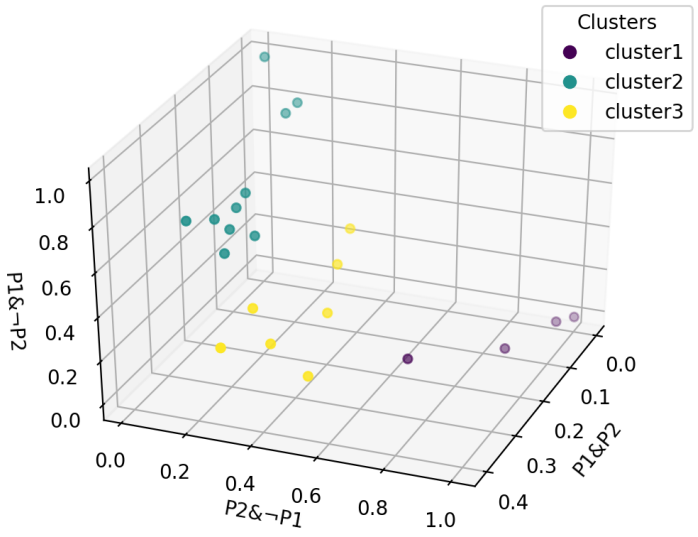


Figure 5.6 3D visualization of the three synchrony classes obtained using the dendrogram.

Relationship between Synchrony and Engagement & Social Attitudes We also want to see if synchrony plays a role in the perception of engagement and social attitudes of warmth and competence. As we have previously hypothesized, recip-

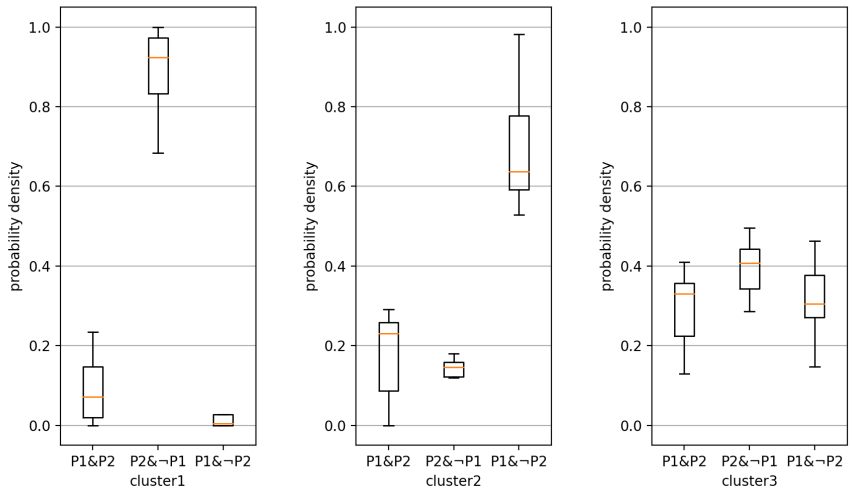


Figure 5.7 Probability density of smiles that are in sync ($P1\&P2$), or not ($P2\&\neg P1$ and $P1\&\neg P2$) for each class obtained with the dendrogram: (left) cluster 1; (middle) cluster 2; (right) cluster 3.

rocal adaptation (synchrony and entrainment loop) being related to engagement and social dimensions of warmth and competence, we hypothesize the following based on previous literature (Biancardi et al. [2021], Lau [1982], Reis et al. [1990], Biancardi et al. [2017], Cuddy et al. [2011]):

- Hypothesis 1: positive correlation between synchrony ($P1\&P2$) and engagement level,
- Hypothesis 2: positive correlation between synchrony ($P1\&P2$) and warmth level,
- Hypothesis 3: negative correlation between synchrony ($P1\&P2$) and competence level.

For the annotations, we base on previous works done on the NoXi corpus (available with the annotation tool NOVA (Heimerl et al. [2019]). For the engagement annotations, the perception of engagement was characterized in Dermouche and Pelachaud [2019a] with five levels (0: strongly disengaged; 1: partially disengaged; 2: neutral; 3: partially engaged; 4: strongly engaged). In Biancardi et al. [2017], the continuous annotations of social dimensions of warmth and competence were done with scores ranging from 0 to 1 (0: very low degree of perceived warmth or competence; 1: very high degree of warmth or competence). As the work of Biancardi et al. [2017] focuses on $P2$ (expert), we also evaluate the impact of synchrony on the three aspects of engagement, warmth, and competence of $P2$.

To test if our assumptions are correct, we observe the engagement and the social attitudes depending on the synchronization levels of 1, 2, and 3 which correspond to the synchrony score $P1\&P2$ of clusters 1, 2, and 3 respectively (obtained from our measures of synchrony). The analysis was done with two different methods.

The first method, method 1, consists of computing the *local average value* of engagement and/or social attitudes levels only on the segments where a smile occurs, either on both participants' faces (condition $P1\&P2$) or for just on one participant's face (condition $P2\&\neg P1$ or $P1\&\neg P2$). A delay of 2 seconds is considered for the reaction lag of the evaluator, as proposed in Marioooryad and Busso [2014]). We then compute the mean of all the averaged values of segments.

The second method, method 2, uses the *global average value* of the engagement level (respectively of the warmth and competence levels) over the entire video of each dyad, independent of the smile synchrony sequence. For this second method, as a single value is computed for each entire video of the corpus, we cannot use it to see the relationship that depends on our measures of synchrony ($P1\&P2$, $P2\&\neg P1$, and $P1\&\neg P2$) as they derive from a single sample (i.e. one smile occurrence).

So all in all, we evaluate the relationship between synchrony and engagement (identically for both social attitudes) using three conditions:

- Condition 1: method 1 and averaged values of segments belonging to ($P1\&P2$, $P2\&\neg P1$, and $P1\&\neg P2$),
- Condition 2: method 1 and averaged values of segments belonging to synchrony levels 1, 2, and 3,
- Condition 3: method 2 for video dyad of synchrony levels 1, 2, and 3.

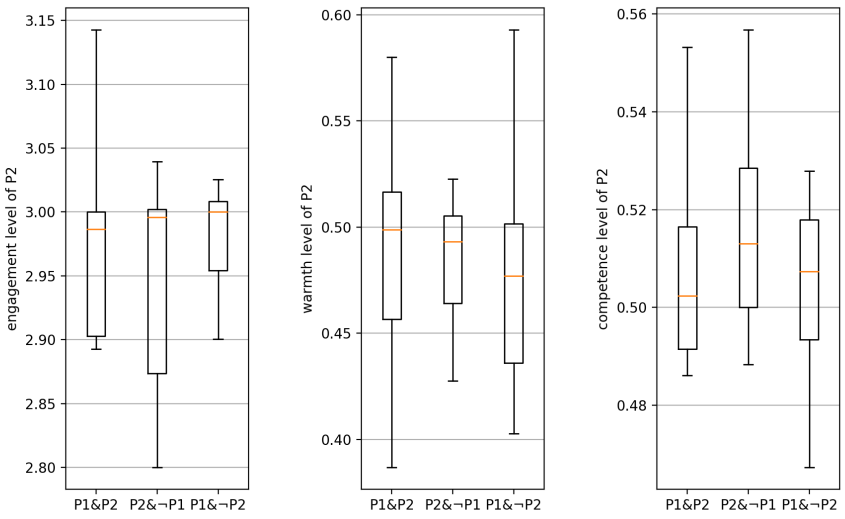


Figure 5.8 Engagement (left), warmth (center), and competence (right) levels measured for condition 1.

For the engagement, we can see in Figure 5.8 (left) that similar levels of engagement are obtained for $P2$ disregarding whether $P1$ and $P2$ are in sync ($P1\&P2 \sim 2.986$) or not ($P2\&\neg P1 \sim 2.996$ and $P1\&\neg P2 \sim 3.0$). When looking at the relationship depending on the synchrony level, in Figure 5.9 (left) we can observe that the level 1 (~ 2.682) indicates a lower engagement level compared to levels 2 and 3 (3.0 for both) and in Figure 5.10 (left) the proportional relationship

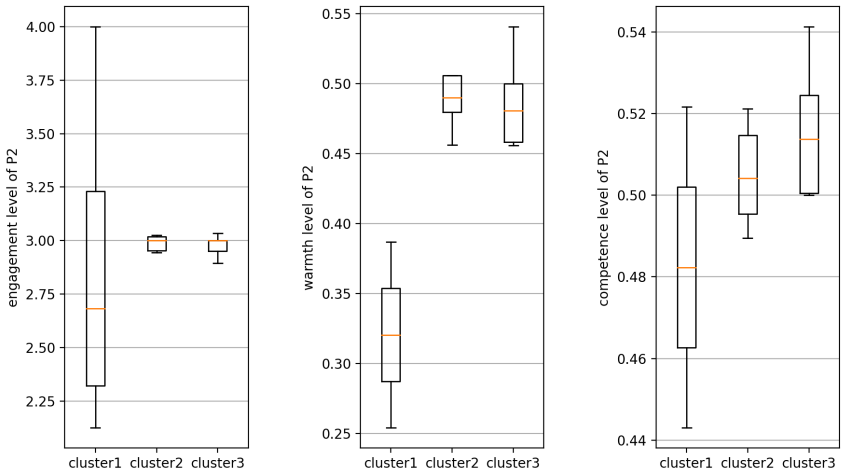


Figure 5.9 Engagement (left), warmth (center), and competence (right) levels measured for condition 2.

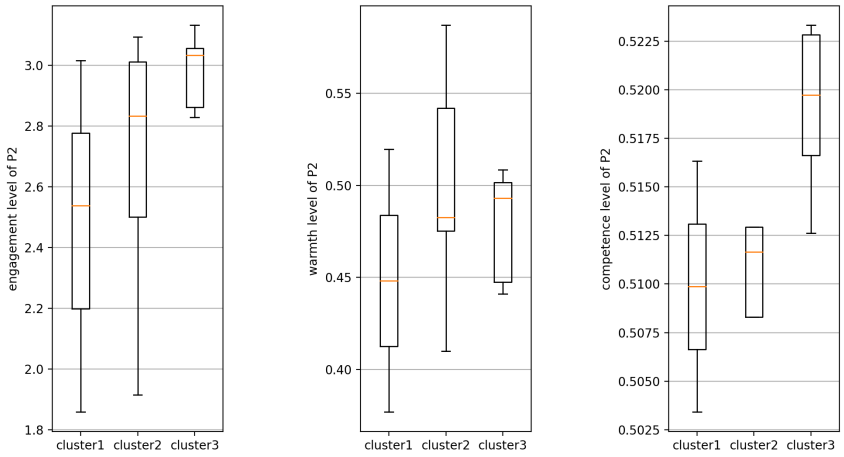


Figure 5.10 Engagement (left), warmth (center), and competence (right) levels measured for condition 3.

between engagement and synchrony level is clearly shown (level 1 ~ 2.538, level 2 ~ 2.832, and level 3 ~ 3.033). Thus, we found a positive relationship between engagement and synchrony levels. Our analysis shows that the more engaged the participants are the more they show behavior synchronization (here smile of $P1\&P2$). It validates our first hypothesis. Condition 3 offers a clearer view. That is, providing a *global average value* for the engagement level better represents the characteristics of engagement of participants in an interaction; only looking at the short sequences of smiling moments is not sufficient to capture the whole picture of the engagement.

The warmth dimension in Figure 5.8 (middle) shows that when $P1$ and $P2$ are in sync (for their smile, at least) $P2$ is perceived warmer ($P1\&P2 \sim 0.499$) compared to when they are not in sync ($P2\&\neg P1 \sim 0.493$ and $P1\&\neg P2 \sim 0.477$). $P2$ is also thought to be warmer when he/she is the only one smiling ($P2\&\neg P1 \sim$

0.493) against the opposite situation ($P1\&\neg P2 \sim 0.477$; only $P1$ smiling). In Figure 5.9 (middle), the lower level of warmth at synchrony level 1 (~ 0.32) is distinguishable from the higher levels of warmth at synchrony levels 2 and 3 (~ 0.49 and ~ 0.481 respectively). When looking at Figure 5.10 (middle), we can see a rise in warmth level as the synchrony level increases (level 1 ~ 0.448, level 2 ~ 0.483, and level 3 ~ 0.493). The results for warmth tell us that being in synchrony with the other interacting participant gives a warmer impression and that the improvement of synchrony level ($P1\&P2$) conducts the growth in warmth level which validates our hypothesis 2. Moreover, the smiling tendency of the interlocutor is linked to his/her impression of warmth which is conformed with the literature that (genuine) smiles are signals of warmth (Lau [1982], Reis et al. [1990]).

In the case of the social trait of competence, we can remark in Figure 5.8 (right) that $P2$ is perceived as more competent when $P2$ is the only one smiling with no smiling back from $P1$ ($P2\&\neg P1 \sim 0.513$) followed up by when $P1$ is smiling alone ($P1\&\neg P2 \sim 0.507$) and then by when $P1$ and $P2$ are in sync ($P1\&P2 \sim 0.502$).

Previous researches (Bernstein et al. [2010], Biancardi et al. [2017]) have highlighted that a smiling person is perceived as more affiliative and less dominant. In the context of an interaction, the interplay of participants' behaviors modulates their perception. In a study on behavior mimicry, Tiedens and Fragale [2003] have reported that when participants have different status (here in NoXi, knowledgeable on a topic vs wanted to learn on this topic), it seems to be correlated with complementarity pattern rather than mimicry. In the NoXi corpus, $P2$ acts as the "expert" that conveys information on a topic that $P1$ is interested to learn more about. Thus, $P2$ has the role of a knowledgeable person on the topic of discussion. It confers him/her a form of expertise and thus of competence. In the context of the NoXi corpus, when $P2$ displays a smile which is not responded by a smile of $P1$, $P2$ appears to be more competent than in the other smiling conditions. However, coordination of behaviors of both participants appears to modulate this inference as reported in previous studies (Tiedens and Fragale [2003]). Further studies involving other nonverbal signals (e.g. frowning, sighting) need to be conducted to see if this condition leads to complementarity.

In Figures 5.9 (right) and 5.10 (right), the increase in synchronization level leads to the rise in the perception of competence level. We could say that the higher the synchronization the more the interlocutors show involvement that gives a feeling of being more proficient around the subject of discussion and thus appearing more competent. This finding is against our hypothesis 3, of synchrony ($P1\&P2$) having an indirect relationship with competence level. Instead it follows previous literature work that saw the phenomenon of smiling people being perceived as intelligent and trustworthy (Lau [1982], Scharlemann et al. [2001]). However, it is against our hypothesis with is based on observation of Biancardi et al. [2017], Cuddy et al. [2011] that smiling behavior is negatively associated with competence. In our case, we remark a halo effect which occurs when the judgments of an undescribed targeted dimension (i.e. competence) goes towards the same direction as the other given dimension (i.e. warmth). Contrary to Biancardi et al. [2017], Cuddy et al. [2011]'s study that looks only at one person, in our study we focus on the interaction and on how participants in a dyad interact

with each other. This could explain the differences in our results and Lau [1982], Scharlemann et al. [2001] and in Biancardi et al. [2017], Cuddy et al. [2011].

5.4 Entrainment Loop measure

5.4.1 Definition

We are also interested in capturing the entrainment of smile. The smile of PA can entrain the smile of PB which then entrains PA to continue to smile or to smile again within a certain time delay and vice versa. We refer to this as the entrainment loop of smile. The entrainment loop consists of two types:

- Type 1: continuous smile, seen in Figure 5.11;
- Type 2: repeated smile with an overlap or within a certain time delay (i.e. mimicry delay of 4 seconds), seen in Figure 5.12 and Figure 5.13 respectively.

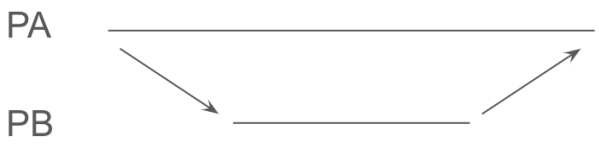


Figure 5.11 Entrainment loop type 1 of a continuous smile of PA.

Figure 5.12 Entrainment loop type 2 of a repeated smile of PA with overlap.

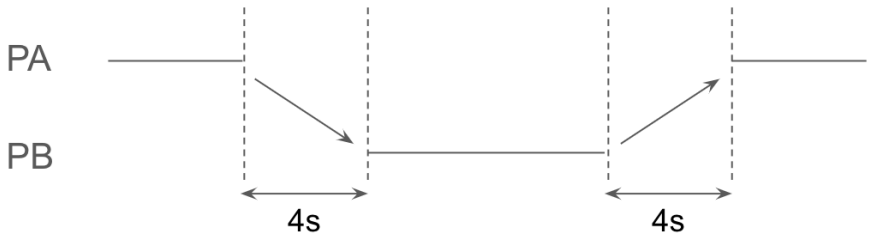


Figure 5.13 Entrainment loop type 2 of a repeated smile of PA within the mimicry delay of 4 seconds.

We capture these two types of entrainment loop and count the number of occurrences of entrainment loops for each interaction.

5.4.2 Analysis

We also want to observe the impact of entrainment loop on the aspects of engagement and social dimensions of warmth and competence.

Types of Entrainment Loop We first check the number of occurrences of the two types of entrainment loop.

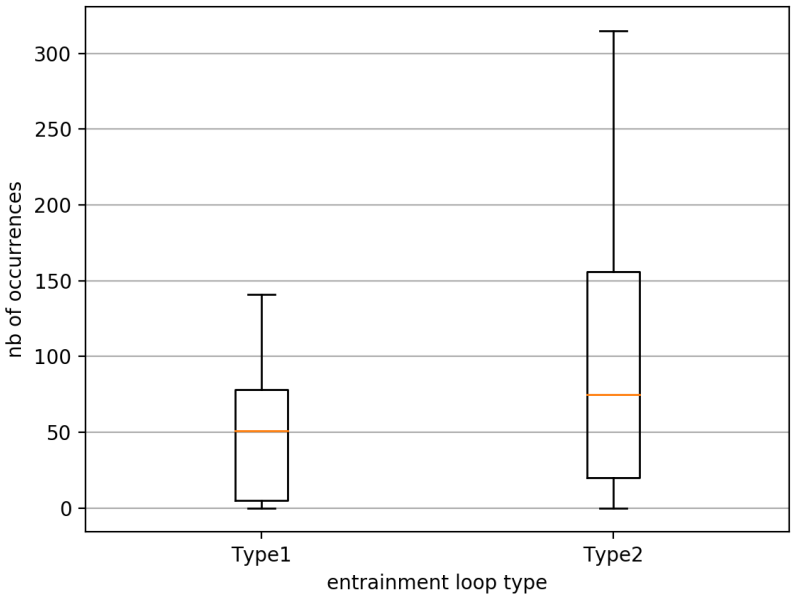


Figure 5.14 Number of occurrences of the two entrainment loop types.

In Figure 5.14, we can notice that the two entrainment loop types' occurrence frequencies are not negligible. With this, we can state that both types should be considered.

Relationship between Entrainment Loop and Engagement & Social Attitudes As above, we observe the relationship of entrainment loop with the aspects of engagement and social attitudes via the aforementioned methods (using method 1: local average value or method 2: global average value of the engagement, warmth, and competence levels). Before analyzing the relationships, we cluster the interactions into two groups by splitting them with the median number of occurrences of entrainment loops (for both types).

For the engagement, we can note that the engagement level increases with respect to the entrainment loop occurrences for both method 1 ($low \sim 2.490$ and $high \sim 2.956$) and method 2 ($low \sim 2.928$ and $high \sim 2.985$), in Figure 5.15 (left) and in Figure 5.16 (left) respectively.

For the social attitudes, when looking at them for method 1, we can remark with their median values that warmth and competence attitude levels decrease ($low \sim 0.493$ and $high \sim 0.480$, and $low \sim 0.512$ and $high \sim 0.507$ respectively),

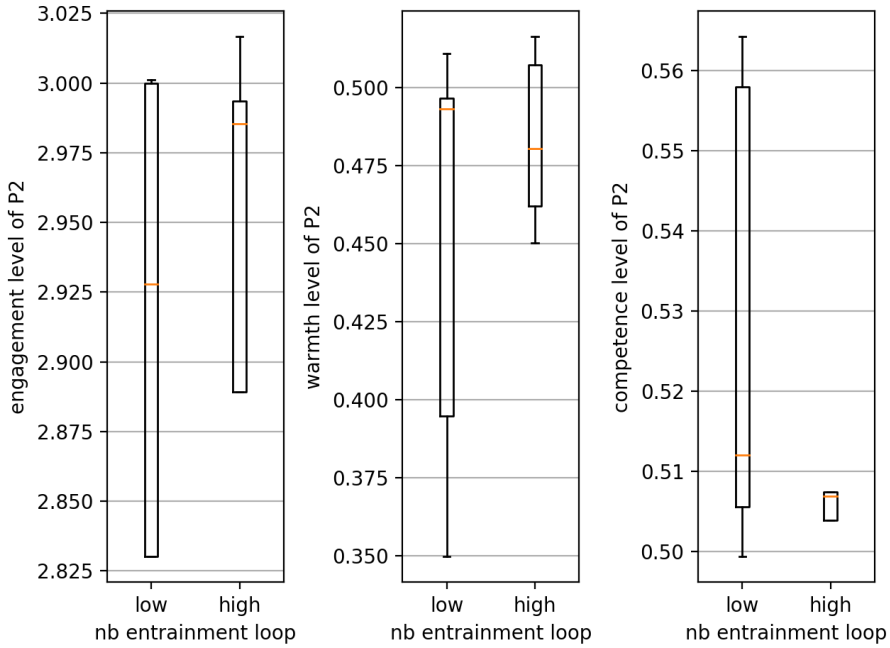


Figure 5.15 Engagement (left), warmth (center), and competence (right) levels measured with method 1 (local average value) for entrainment loop.

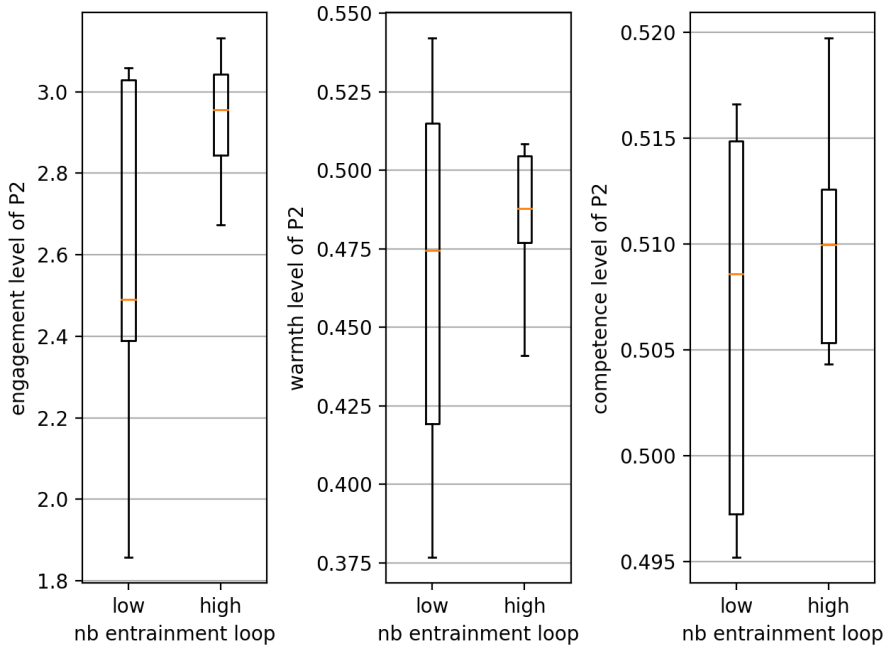


Figure 5.16 Engagement (left), warmth (center), and competence (right) levels measured with method 2 (global average value) for entrainment loop.

in Figure 5.15 (left) and (right) respectively, when entrainment loop occurrence transits from low to high. Nevertheless, for both cases, we can see that the class for high entrainment loop occurrence is more concentrated ranging at a high warmth

level ($0.394 < low < 0.497$ and $0.462 < high < 0.508$) and low competence level ($0.505 < low < 0.558$ and $0.503 < high < 0.508$) in Figure 5.16 (left) and (right) respectively. Thus, we can state that at the moment of the entrainment, the warmth level rises and the competence level decreases which is inline with the findings of Biancardi et al. [2017], Cuddy et al. [2011].

For method 2, both warmth and competence levels increase ($low \sim 0.475$ and $high \sim 0.488$, and $low \sim 0.509$ and $high \sim 0.510$ respectively). However, no significance can be found for competence, thus validating only for warmth level to be correlated to the number of entrainment loops.

5.5 Discussion

As *reciprocal adaptation* occurs naturally as we converse, it generally passes unnoticed without giving any explicit attention to it. Nevertheless, this aspect of *reciprocal adaptation*, particularly interpersonal synchronization and entrainment loop, is an important factor for interactive and engaging communication. With our new *reciprocal adaptation evaluation measures*, that assess synchrony behaviors including their response absences and measures entrainment loop, we were able to carry out several statistical analyses on smile synchrony, clustering synchronization levels (level 1, level 2, and level 3), and the relationship with engagement and social dimensions (warmth and competence). Also, we observed the relation between entrainment loop occurrences and engagement and social dimensions.

We validated our hypotheses of observing a positive correlation between synchrony and entrainment loop with engagement and warmth, while we see a halo effect for competence. Thus, we can say that *reciprocal adaptation,* which is assessed via our measures, also has a direct relation with engagement and social attitude.

Behavioral expressions such as other facial muscle movements (e.g. eyebrow and mouth), hand/body gestures might also be relevant to model synchrony and entertainment. They may require a more complex modeling but it will be interesting to study them and to check that they also show similar relationships with the dimensions of engagement and social attitude.

Our *reciprocal adaptation measures* (three synchrony measures and entrainment loop measure) can be used to evaluate if the agent produced human-like behaviors with reciprocal adaptation for human-agent interaction. This can be done by comparing the values obtained by the human-agent interaction against those obtained from human-human interaction. To detail, the human-agent interaction quality can be assessed by checking if the results of the agent, obtained via our *reciprocal adaptation measures*, show similar behaviors with those of the real human-human interaction for both synchrony behaviors including their absence of response and behavior entrainment loop.

5.6 Contributions and Conclusion

5.6.1 Contributions

Our contributions are as follows:

- We propose new *reciprocal adaptation* measures.
- We observe a direct relationship between *reciprocal adaptation* (interpersonal synchrony and entrainment loop) and dimensions of engagement and warmth.

5.6.2 Conclusion

In this chapter, we propose novel measures of *reciprocal adaptation* which can be used as objective measures to evaluate the agent behavior quality in human-agent interaction. In detail, they can be used to measure whether an agent (i.e. agent behavior generating computational model) offers similar adaptation properties as a human in an interaction. Through these measures, we were able to find a positive relation between reciprocal adaptation and the dimensions of engagement and social attitude. These results are promising for the field of human-agent interaction in providing a new way of evaluating the interaction quality.

The key points of this Chapter:

Addressing Hypothsis

- *Reciprocal adaptation* (synchrony and entrainment loop) is related to engagement and social dimensions of warmth and competence of the interlocutors.

Reciprocal adaptation measures

- New objective measures for the evaluation of the agent behavior quality in human-agent interaction.
- Assess synchrony behaviors including their response absences and measure entrainment loop.
- *Reciprocal adaptation* has a direct relation with engagement and warmth.

Publication

- Jieyeon Woo, Catherine Pelachaud, and Catherine Achard. Reciprocal adaptation measures for human-agent interaction evaluation. In *ICAART*, 2023e

Part III

Reciprocally adaptive SIA

Chapter 6

Reciprocally adaptive SIA behavior generation

Contents

In this Chapter, we present the modeling of *reciprocal adaptation* concentrating on the modelization of *interpersonal temporality*, *multimodality*, and behavior *continuity*. Our proposed Augmented Self-Attention Pruning (*ASAP*) model renders *natural* and *human-like* behaviors, as both *listener* and *speaker*, that are *engaging* and *in sync* with the interlocutor.

6.1 Introduction

We aim to provide *SIA*s with the capacity of *reciprocal adaptation* to enhance their behaviors so that they can behave naturally like a human. We use *multimodal* features (visual and acoustic) and produce *SIA* behaviors of an active interactant as both *listener* and *speaker*. We hold attention to the aspect of behavior *coherence*, *synchrony*, and *continuity*. Behaviors are made up of continuous values that evolve over time (for example for human motion the body landmark positions change smoothly in time). We also intend to assure the production of continuous behaviors by looking at their *temporal continuity*. Behavior motions should not only be *continuous* but also *coherent* and *in sync* with those shown by the interactant. We thus focus on the *temporal alignment* and the *appropriateness* of the generated *SIA* behavior type (e.g. a smile in response of an interlocutor's smile). Along with modeling of such properties, we look into how the *quality* of the generated behaviors could be quantified via objective measures.

For this study, we pose the following research questions (RQs):

- RQ 1: endowing *reciprocal adaptation* capability to the *SIA* improves the *interpersonal dynamics* (synchrony and engagement) of the generated agent behaviors;
- RQ 2: agent behavior *quality* (naturalness and human-likeliness) can be enhanced with the modeling of *reciprocal adaptation*.

To create such a *SIA* that adapts its behaviors to its interlocutor, we propose the Augmented Self-Attention Pruning (*ASAP*) model. *ASAP* models the *reciprocal adaptation* of interaction partners throughout the interaction using multimodal signal information of both interlocutors along with the *interpersonal relationship* between them.

This chapter is structured as follows: Section 6.2 presents a brief state of the art of related techniques for continuous nonverbal behavior prediction and evaluation measures; Section 6.3 introduces the problem that is being addressed; Section 6.4 details the implementation of our *ASAP* model and provides objective and subjective evaluation results; Section 6.5 summarizes our findings.

6.2 Related Works and Limitations

We are interested in generating *social* nonverbal behaviors of *SIA*s. The gesture generation task is similar to sequence forecasting in the sense that we are predicting the future sequence depending on past information. This motivates us to learn from techniques that forecast sequences and apply them to our goal of generating agent nonverbal behaviors. Sequence prediction consists of two types which are offline prediction and online prediction.

Offline prediction infers the entire sequence at once. One of the most popular uses of this technique is sequence to sequence (Seq2seq) prediction. Seq2seq models have shown promising results for various applications such as machine translation (Sutskever et al. [2014], Moslem et al. [2023]) and speech recognition (Li et al. [2018], Radford et al. [2023]).

Online prediction, on the other hand, outputs sequentially making predictions for each time-step individually. This prediction technique is popularly used for time series forecasting (Kumar and Jha [2013], Wan et al. [2019], Lippi et al. [2013], Tian and Pan [2015], Kim [2003], Tsantekidis et al. [2017]) which can be grouped into two types: sliding window prediction (input data inside a sliding window are used to predict the next time-step) and adaptive online prediction (update of the latent vector using the full input data sequence rendering *continuous* output data). To further enhance output *continuity*, observations from previous time-steps can be used. The technique of feeding back the output to the model for the next prediction is referred to be autoregressive. This can be applied to both sliding window and adaptive online predictions.

The memory retention present within recurrent networks such as RNN, LSTM, and TCN, has shown great promise in time series forecasting. As human behaviors heavily depend on previously performed ones, this aspect of memory is also important for our situation. Moreover, as behavior must be *continuous*, employing the adaptive online prediction in an autoregressive manner is preferable.

For our study, we focus on dyadic interactions which leads us to concentrate on modeling the *temporal* relationship between participants during an interaction. We look into the literature that models *interpersonal* relationship (or *reciprocal adaptation*).

In the work of Feng et al. [2017] and Dermouche and Pelachaud [2019b], *SIA*'s *facial gestures* are synthesized based on past gestures of both, *SIA* and *User*, without considering the existing relation with *audio modality* (Pell [2005], Yehia et al. [2002]), and do not ensure the *motion continuity*. Grafsgaard et al. [2018] synthesize interlocutor's *gestures* based on the interlocutors' *audio* and their *facial modalities*. In the work of Jonell et al. [2020], *SIA*'s *facial gestures* are generated based on *SIA*'s *speech* and the *User*'s *speech* and *facial gestures*. However, these models (Grafsgaard et al. [2018], Jonell et al. [2020]) are prone to produce *non-continuous gestures*. Ahuja et al. [2019] generates body pose sequences based on the *User*'s audio and body pose and audio of the *SIA* by capturing interpersonal and intrapersonal dynamics via selective attention. Tuyen and Celiktutan [2022] predicts the upper body motion (face, body, and hand landmarks) with a context-aware encoder-decoder model learning from contextual information (encoding of interacting partner's nonverbal behaviors of body motion and audio) and *SIA*'s body motion. The works presented in Ng et al. [2022] ensure *SIA*'s *behavior continuity* by employing autoregressive online inference while modeling *SIA*'s and *User*'s multimodal features. Only the *listener*'s behavior is modeled.

The aforesaid works for behavior generation model the *interpersonal* relationship and/or the *multimodal* signals. The multimodal aspect is missing in Feng et al. [2017], Dermouche and Pelachaud [2019b] and Feng et al. [2017], Grafsgaard et al. [2018], Dermouche and Pelachaud [2019b], Ahuja et al. [2019], Jonell et al. [2020], Tuyen and Celiktutan [2022] do not assure the continuity aspect. Ng et al. [2022] ensure both aspects but their model cannot be employed with a small training data. We try to resolve these issues and model the *reciprocal adaptation* by capturing the *interpersonal temporality* and *multimodality* while ensuring the generated agent's behavior *continuity* and the functioning with a small amount of training data.

6.3 Problem Definition

In this chapter, we focus on generating nonverbal behavior (*SIA*'s facial expressions and head/gaze movements; ref. Chapter 4) for dyadic interactions. The rendered behaviors should be *continuous* and capture *interpersonal relationship*. The agent should be able to produce behavior for both *speaker* and *listener* roles and the behavior prediction model should ideally work even with a small database.

With the goal to create a *SIA* capable of adapting its behaviors to its interlocutor, we propose the Augmented Self-Attention Pruning (*ASAP*) model that models the *reciprocal adaptation* of interaction partners throughout the interaction. The *multimodal* signal information of both interacting partners along with the *interpersonal relationship* between them are captured. Specifically, *ASAP* allows us to: (1) capture *multimodal* information of visual and acoustic features; (2) learn from both interactants through data augmentation technique; (3) better select key features within the interaction via the self-attention mechanism with pruning; (4) generate continuous nonverbal behaviors by updating cells' memories at each step of the inference phase with autoregressive adaptive online prediction; (5) generate behaviors as both active *listener* and *speaker*; (6) and train without needing a massive amount of data.

6.4 Augmented Self-Attention Pruning (ASAP) Model

We hold interest in generating *social* nonverbal behavior of a *SIA* (be a *speaker* or a *listener*) when interacting with its human interlocutor. In particular, we aim to model the *reciprocal adaptation*, by capturing the behavior coordination of both interactants, notably the *interpersonal relationship*.

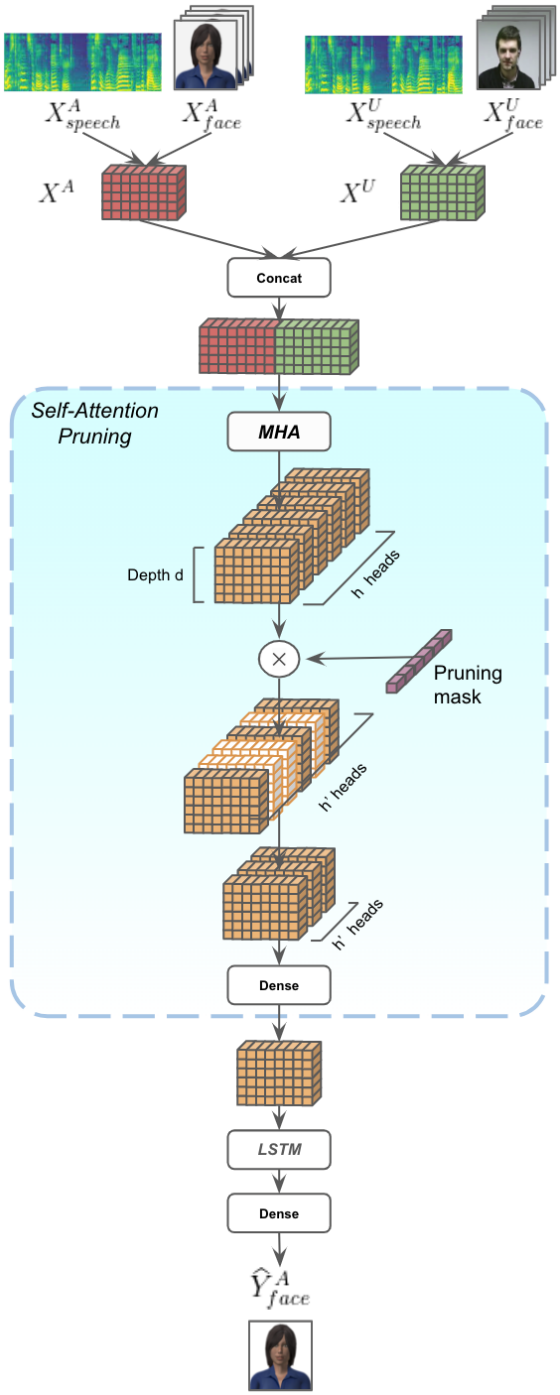
6.4.1 Model Architecture

We propose a new architecture that models the *reciprocal adaptation* which is our Augmented Self-Attention Pruning (*ASAP*) model[1][2], as illustrated in Figure 6.2. It takes 100 previous frames $(t-99:t)$ for both human user and *SIA* to predict the agent behavior of the next frame $(t+1)$. *ASAP* consists of three key techniques: data augmentation technique, self-attention pruning, and autoregressive adaptive online prediction. Each technique has its own usage which are as follows. The data augmentation technique learns from both interactants enabling the training without needing a massive amount of data. The self-attention pruning selects key features within the interaction from multimodal information. The autoregressive adaptive online prediction generates continuous *SIA* nonverbal behaviors.

Data augmentation

Since our database is not that large, we make use of a data augmentation technique. To learn the *reciprocal adaptation* we need accurate data from both partic-

[1]The code is available here: https://github.com/jieywoo/ASAP.
[2]The demonstration video is available here: https://www.youtube.com/watch?v=feojlOrFCIg.



ASAP model architecture

Figure 6.2 ***ASAP*** (**A**ugmented **S**elf-**A**ttention **P**runing) model architecture. The self-attention pruning section takes the *speech* X_{speech} and the *facial gestures* X_{face} of the previous 100 frames of both the *SIA* (A) and the *User* (U) to learn the *interpersonal relationship* (or *reciprocal adaptation*) between them. The *SIA*'s *facial gesture* for the next frame at $t+1$ $\widehat{Y}^A_{face}$ is generated. To infer the next A's behavior, we feed back the predicted A's behavior and the ground truth of U.

ipants. This leads us to propose a data augmentation technique that learns from both interlocutors in an equal manner, during the training phase, instead of using classical data augmentation techniques, such as adding noise or dropouts. That is, we learn from the characteristics of both interacting partners. For each batch of the training phase, we assign randomly the interlocutor identity that will be played by the agent to one of the interlocutors. We learn to predict the behaviors of this interlocutor. Then, we follow by alternating and assigning the interlocutor identity for the agent to the other interlocutor and continue the learning process. For a better understanding, we refer to each interacting person of a dyad as $U1$ for person 1 and $U2$ for person 2. There are two possible choices of giving the *SIA* the interlocutor identity of either $U1$ or $U2$. During each batch, the interlocutor identity of the *SIA* is reassigned randomly (to either maintain the same identity of the previous batch or to switch identities from $U1$ to $U2$ or $U2$ to $U1$). The *SIA* learns to generate the behavior of the corresponding interlocutor identity. The data augmentation simulates the interlocutors' behaviors without separating whether it's those of a *speaker* or a *listener*. It only takes into account the interlocutor identity (either $U1$ or $U2$). By doing so, the model learns to predict equally the behaviors of both participants and focuses on modeling the interaction between the two rather than the specific characteristics of a single person. At the inference stage, the *SIA* will be one of the interlocutors ($U1$ or $U2$) and the user will be the other interlocutor ($U2$ or $U1$ reciprocally). Thus, the data augmentation is done with the identities of the *SIA* A and the user U instead of $U1$ and $U2$.

Self-Attention Pruning

To better model the *reciprocal adaptation*, we want to capture *interpersonal relationship* (interpersonal behavior coherence and synchrony) and *multimodality* from key features. The selection of relevant features is done via an attention mechanism. A self-attention, using the multi-head attention of the Transformers (Vaswani et al. [2017]), is performed using all interlocutors and visual and acoustic modalities (X^A_{speech}, X^A_{face}, X^U_{speech}, and X^U_{face}). The self-attention layer captures key information to model which behaviors should occur along with mimicry and synchronization mechanisms all at once. However, most attention heads within the multi-head attention (MHA) contain redundant information (Michel et al. [2019], Voita et al. [2019]) which leads the model to overfit. Michel et al. [2019] and Voita et al. [2019] demonstrate the overfitting problem caused by redundant attention heads can be solved by applying pruning (i.e. pruning removes redundant heads). Our aim is to modelize the *reciprocal adaptation*, by retrieving key information via pruning. Pruning allows us to drop repetitive heads only rendering attention to dissimilar heads encoded with unique information and it also increases the inference speed. The pruning of attention heads is similar to structured pruning where neurons are pruned. An example of structured pruning is given in Figure 6.3.

Instead of pruning the neurons, we prune the attention heads. Our technique differs from the conventional pruning technique which prunes a given percentage of less significant neurons or connections (for unstructured pruning). Once the model is trained, the same neurons/connections are pruned out disregarding the

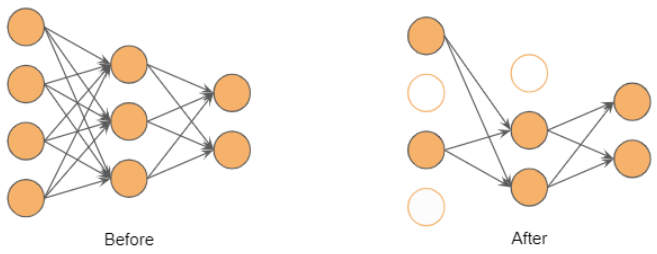


Figure 6.3 Example of structured pruning.

input. For our pruning technique, we learn to choose which head(s) are meaningful for each specific frame via a pruning mask. For each input sequence, a custom pruning mask is applied. To detail, as seen in Figure 6.2, the input sequence that consists of $T = 100$ frames from $t-T+1$ to t are passed through the MHA (with the depth of d and h attention heads). A custom pruning mask is learned to minimize the loss of the network for each input sequence to prune the attention heads (each with the dimension of $d \times T$). The custom pruning mask selects to learn from a certain number of attention heads h' out of h heads. In this way, the pruned attention heads vary for each prediction. For each head, the significance factor is obtained by applying a sigmoid function $\sigma(h) = \frac{1}{1+e^{-h}}$ element-wise and then binarized (by rounding) within the pruning mask. To detail, the pruning mask is a vector of dimension h where each element corresponds to the significance factor of each h MHA head which is obtained via the sigmoid function. The significance factor is binarized for each element to only leave significant heads as 1 and the rest as 0. Non-significant heads are removed after applying the pruning mask to the attention heads outputted by the MHA. Then, the information of the key heads is grouped together (dimension reduced from $h \times d \times T$ to $h' \times d \times T$), and then the essential information among the information of the key heads is obtained via a dense layer (or fully connected layer; the self-attention pruning module rendering the final dimension of $d \times T$). With our pruning technique, we can ensure that our model accesses only unique and relevant information for each prediction.

Autoregressive adaptive online prediction

We want to generate *continuous SIA* behaviors $\widehat{Y}^{A}_{face}$ which are assured by applying the adaptive online prediction. During the whole course of the interaction, the model updates the LSTM memory in a *continuous* way as in Yang et al. [2017]. Non-continuous values come from the predictions that are made independently for each input sequence without conserving previous memories (i.e. temporal sliding window). By applying adaptive online prediction during the inference phase, we circumvent this problem as the past information is kept within the memory cells of the LSTM and used to make new predictions. Also, the prediction is made in an autoregressive fashion by feeding back the predicted values of previous time-steps as input for the prediction at the next time-step.

6.4.2 Implementation Details Training Regime

All models are implemented in Tensorflow and trained using the French *NoXi* dataset (Cafaro et al. [2017]) (ref. Chapter 5). They are trained for 1000 epochs on $2.20GHz$ Intel Xeon Linux server with NVIDIA GeForce GTX TITAN X and $64GB$ RAM. They all share the same parameters: batch size of 32 and Adam optimizer (Kingma and Ba [2014]) with a linear learning rate scheduler (learning rate starting from 0.001, factor 0.2 decay on plateau, and patience 3). The best set of hyperparameters is chosen for each model after manual optimization, via manual grid search, based on the validation set. For *ASAP*, after fine-tuning the MHA, four attention heads ($h = 4$) with a depth of $d = 16$ and cell sizes of $c_{SelfAP} = 64$ for the dense layer within the self-attention pruning technique, $c_{lstm} = 20$ for the LSTM layer, and $c_{dense} = 20$ for the final dense layer were used. Mean Squared Error (MSE) was used as the objective function during the learning stage. The dataset is split for training:validation:testing in the ratio of 70:10:20 and we ensured that the test set contains pairs of dyads that were never seen in the train and validation sets. To assure that the training and test sets do not include the same person, we have manually excluded participant pairs for the test set.

6.4.3 Database and Feature Extraction

The French NoXi database (Cafaro et al. [2017]), introduced in Chapter 4, is used for the *ASAP* model training and evaluation. *ASAP* employs visual features of eye movements (G_x and G_y), head rotations (R_x, R_y, and R_z), 6 upper face AUs ($AU1$, $AU2$, $AU4$, $AU5$, $AU6$, and $AU7$), and smile ($AU12$). ***Speech and facial gestures*** are highly tied together (Pell [2005], Yehia et al. [2002]). With this relation, we focus on capturing visual prosody, that is we use *speech* information to drive *SIA*'s *facial gestures*. We utilize audio features of fundamental frequency, loudness, voicing probability, and 13 MFCCs. As such the dimensions of X_{speech} is $Tx16$ (X^A_{speech} or X^U_{speech}), X_{face} is $Tx12$, and $\widehat{Y}^A_{face}$ is $1x16$.

6.4.4 Objective Evaluation

Our goal is to evaluate if *ASAP* captures the *reciprocal adaptation* between two participants, that is the *interpersonal relationship* encoded with multimodal signals. Also, we check the quality of our generated *SIA* behavior with both roles as *listener* and *speaker*. We compare the performance of *ASAP* to that of two recent state-of-the-art models, which are the works of Dermouche and Pelachaud [2019b] and Woo et al. [2021], by evaluating their generated nonverbal behaviors both quantitatively and qualitatively.

As mentioned above in Section 6.2, evaluating nonverbal behaviors has always been a challenge. Until now there is no perfect measure that can thoroughly quantify the dynamics of the behaviors. To assess our model, we propose to use several objective measures, one metric for each measuring type (i.e. point to point, statistical, and resemblance).

Objective Evaluation Measures

As point to point measure, we use the RMSE (Feng et al. [2017], Dermouche and Pelachaud [2019b], Ng et al. [2022]), to evaluate our generated nonverbal behaviors. This measure provides information on the quality of learning. However, it is not always pertinent to compute the exact behaviors that may arise during an interaction, as different reactions (behaviors) of a participant may arise. Indeed, it is difficult to reproduce the same behavior of a person from a database that contains various participants (excluding the targeted person) each possessing a personality and showing different behaviors. We chose to use another measure to further evaluate our model.

We are interested in measuring if the behaviors generated by our model have similar distributions as in the NoXi database. That is we check if both, predicted behaviors and ground truth have a similar number of occurrences. Taking the smile as an example, during the course of a conversation the smile intensity of a participant varies continuously. In the NoXi database, the intensity distribution of smiles is more concentrated around subtle and low levels (with a percentage of 84%). We want to assess the quality of the produced nonverbal behaviors globally not on the sequence level but on the entire interaction. Using the example of smiles, we want to see if smiles are predicted throughout the interaction in terms of the distribution of smile intensity level. For this purpose, we check the probability distribution similarity using statistical measures. As previously presented measures for behavior likelihood, in Chapter 3, do not suit our case, we propose the usage of Kolmogorov-Smirnov (KS) two-sample test (Massey Jr [1951]). Its use is new to behavior quality evaluation. The KS test is a statistical measure that estimates the quality in a quantitative manner by measuring the difference in density probability between the ground truth and the generated sequence for each output dimension. The KS test measures the distance between the generated $\hat{y}(t)$ and real $y(t)$ data distributions (or more precisely the cumulative distribution functions $F_{\hat{y}}(t)$ and $F_y(t)$):

$$KS_{dist} = \max_t |F_{\hat{y}}(t) - F_y(t)| \tag{6.1}$$

The KS test is applied independently for each feature and the average score is calculated.

Point to point metrics and statistical measures for density distribution do not capture the temporal dependencies that exist between partners. To better observe the temporal dependencies, we employ the Dynamic Time Warping (DTW) (Müller [2007]). DTW measures the similarity between two temporal sequences that may vary in speed and length. DTW, like the RMSE, can be used between $\widehat{Y}^A(t)\&Y^A(t)$, where $\widehat{Y}^A(t)$ is the generated agent's behavior and $Y^A(t)$ is the human ground truth behavior representing the agent. Instead of having another precision measure, we want to measure whether the *reciprocal adaptation* is well captured. The presence of *reciprocal adaptation* (interpersonal temporal dependency) is verified by seeing if the interlocutors show similar behaviors, responding to each other. We check the proximity (resemblance) of the generated agent's behavior and that of the interacting human ($\widehat{Y}^A(t)\&Y^U(t)$) and the proximity of the behaviors between

both humans ($Y^A(t)\&Y^U(t)$) to see if the agent behavior shows the same adaptation trends as seen in the ground truth. The DTW distance does not have to be small. Actually, it would be easy to copy the behavior of the human at the previous moment to have a DTW of almost zero. This high resemblance between partners can be perceived as an everlasting imitation (like a parrot) and thus may rather hinder the perception of human-like behavior. Thus, DTW between $\widehat{Y}^A(t)\&Y^U(t)$ must be similar to the DTW between $Y^A(t)\&Y^U(t)$ and not necessary small. As our interactions are very long (around $20min$ for each interaction), we compute the DTW in small chunks of $1min$ and a stride of $30s$. Applying DTW in chunks speeds up the computation. All the chunks cover the whole interaction.

Smile (*AU12*), a key socio-emotional signal (Knapp et al. [2013]) frequently observed during an interaction, is produced by both speaker and listener and is often imitated between interlocutors (Hess and Bourgeois [2010]). Previous studies have demonstrated that smile helps *SIA*s to better manage their interaction with their human users (Wang and Ruiz [2021], Ochs and Pelachaud [2013]). Thus, for DTW distance evaluation of $\widehat{Y}^A(t)\&Y^U(t)$, we focus on the smile.

Objective Evaluation Results and Discussion

To compare our model with that of the literature, we need to use the same features. As a result, we first evaluate our model with the features presented in Dermouche and Pelachaud [2019b] (features set 1) and then with those in Woo et al. [2021] (features set 2). The features set are composed as the following:

- Features set 1: only visual features (eyes movement, head rotation, and $AU12$ intensity and activation) of both interlocutors along with conversational state (ref. Chapter 4) inputted to predict visual features of the *SIA* at $5fps$;
- Features set 2: visual and acoustic features (eye movement, head rotation, upper face AUs and $AU12$ intensities, fundamental frequency, loudness, voicing probability, and 13 MFCCs) of both interlocutors to predict the visual features (including upper face AUs) of the *SIA* at $25fps$.

Concerning the evaluation of the eye movement, we evaluate the value of the eye angles like we do for the head rotation. However, we cannot assess if the predicted eye movement corresponds to looking at the same target (e.g. its interlocutor) as in the ground truth as this information is not available in the NoXi dataset (both cameras recording the two interlocutors are not calibrated).

All models were trained and their behaviors were generated for each features set. We conduct an objective evaluation for the two sets of features.

The performance of *ASAP* is compared with the baseline models for each features set using the proposed objective evaluation measures. We consider the following *baseline models*:

- **IL-LSTM** (Dermouche and Pelachaud [2019b]): models only the *interpersonal* relationship based only the visual modality (facial gestures) of both the agent and human user,

Methods		RMSE	KS test
Features set 1	IL-LSTM	0.172	0.298
	sym-IL-LSTM	0.171	0.293
	ASAP (ours)	**0.131**	**0.115**
Features set 2	IL-LSTM	0.444	0.559
	sym-IL-LSTM	0.374	0.415
	ASAP (ours)	**0.239**	**0.301**

Table 6.1 Average RMSE and KS test results for features set 1 and 2.

Features set	Method	DTW $Y^A(t)\&Y^U(t)$ (Ground truth)	DTW $\widehat{Y}^A(t)\&Y^U(t)$
Features set 1	IL-LSTM	21.7	27.3
	sym-IL-LSTM		27.2
	ASAP (ours)		**26.9**
Features set 2	IL-LSTM	1317.5	1562.7
	sym-IL-LSTM		257.5
	ASAP (ours)		**1399.3**

Table 6.2 DTW of smile for features set 1 and 2.

- **Symmetrized IL-LSTM with online LSTM (sym-IL-LSTM**; Woo et al. [2021] **)**: models the *interpersonal* relationship based on *multimodal* features (speech and facial gestures) of both the agent and human user, and assure *motion continuity*.

In Table 6.1, the three models of each features set are evaluated quantitatively by computing the RMSE and performing the KS two-sample test. The KS test was used as it statistically measures the probability distribution similarity between our predictions and ground truth (real interaction). The average score of the output features is calculated (average of 6 output features scores (2 eyes angles, 3 head rotations, and $AU12$ intensity) for features set 1 and that of 12 output features scores (2 eyes angles, 3 head rotations, and the intensities of 6 upper face AUs and $AU12$) for features set 2). From both features sets 1 and 2, we can observe that the RMSE and the KS test scores have better values for *ASAP* than the baseline models.

The DTW between $Y^A(t)\&Y^U(t)$ represents the distance (resemblance) between the signals of the two human participants' interlocutor identities of $U1$ and $U2$ ($U1$ being the human representing the agent and $U2$ being the human interlocutor). The DTW distance is interpreted as the closer the distance gets, the more the two signals of $Y^A(t)$ and $Y^U(t)$ are similar. We check if the models' DTW distance $\widehat{Y}^A(t)\&Y^U(t)$ is close to that of the ground truth interaction (human-human interaction) $Y^A(t)\&Y^U(t)$.

As stated above, smile is a key social signal that is apparent to improve *SIA*'s interaction which leads us to focus on smile. We can see, in Table 6.2, that for smile of features set 1, our *ASAP* performs better than the baseline models in terms of having the DTW distance the closest to the ground truth DTW (26.9, 21.7 respectively). The same conclusion can be drawn for features set 2 (1399.3, 1317.5 respectively). Note that the small value obtained with sym-IL-LSTM model can be

interpreted as a close imitation of the behavior of its interlocutor that may deter the perception of the behavior to be human-like.

Therefore, we can conclude that our *ASAP* model outperforms the baseline models for the three objective evaluation methods, that is RMSE, KS test, and resemblance via DTW distance between $Y^A(t)\&Y^U(t)$.

6.4.5 Subjective Evaluation

Relying only on objective evaluations is not enough to fully assess the quality of the generated agent's behavior. We perform a user perceptive study to complement the objective evaluation where we look more particularly at how the generated multimodal signals and the modeling of the *reciprocal adaptation* (*interpersonal relationship*) by our model influence: 1) the perception of the generated agent behaviors' naturalness and human-likeliness; 2) the perception of the *interpersonal* dynamics such as the synchrony between the interlocutors and the perception of their engagement. To evaluate these aspects of human-agent interaction, we ask the participants to score the interacting *SIA* along 4 measurement constructs: behavior naturalness, behavior human-likeliness, interaction synchrony, and engagement.

Questionnaires to evaluate the perception of behavior naturalness, behavior human-likeliness, and engagement are formulated based on existing questionnaires of human-agent interaction evaluation (Fitrianie et al. [2020], Von der Pütten et al. [2010]). We use a set of three synonyms and antonyms for each dimension. To evaluate the perception of synchrony, we use the dyadic stances of mutual understanding, attention, agreement, interest, and pleasantness proposed by Prepin et al. [2013], Louwerse et al. [2012].

A set of 14 questions (3 for each construct of behavior naturalness, behavior human-likeliness, and engagement, and 5 for interaction synchrony), listed in Table 6.3 are used. The users are asked to answer each question using a Likert scale of 5 points (ranging from 1 (strongly disagree), 2 (disagree), 3 (neutral), 4 (agree), to 5 (strongly agree)). The user's answers are grouped into the 4 constructs by averaging their values.

Subjective Evaluation Method

The evaluation is done via Prolific, an online crowd-sourcing platform. 20 video clips of an approximate duration of $7s$ are manually extracted from the human-human videos of NoXi. In each video clip, a human participant has the speaking turn (talking about a common subject) or is the listener (expressing nonverbal behaviors with visual and acoustic feedbacks which include backchannels such as "ok" and "yes") and the other human participant is either, respectively, *listener* or *speaker*.

For our study, we compare four conditions (the three models which are: *ASAP* and our two baseline models of IL-LSTM and sym-IL-LSTM) with the features set 2 and the ground truth human-human interaction from NoXi (GT)). To evaluate the quality of these conditions, we replace one of the human participants (being the *speaker* in 10 video clips and *listener* in the other 10 video clips) with a *SIA* whose

Construct	Question
Naturalness	The behavior of the virtual character is artificial.
	The behavior of the virtual character is realistic.
	The behavior of the virtual character could exist in reality.
Human-likeliness	The virtual character behaves like a human.
	The behavior of the virtual character resembles a machine.
	The behavior of the virtual character resembles that of a human.
Engagement	The virtual character is engaged in the conversation.
	The virtual character pays attention to the human speaker.
	The virtual character ignores the human speaker.
Synchrony	The virtual character and the human understands each other.
	The virtual character and the human agree with each other.
	The virtual character and the human pay attention to each other.
	The virtual character and the human are interested in the discussion.
	The virtual character and the human have a good time together.

Table 6.3 Set of 14 questions used for subjective evaluation.

behavior is driven by the computational models or the GT. The *SIA* was animated using the open source Greta *SIA* platform (Niewiadomski et al. [2009]) by passing visual features (predictions of the computational models or the GT) along with the audio of the GT. An image of a video is shown in Figure 8.2 in which it displays a *SIA* (left side of the screen) and a human participant (right side of the screen). The lower face was blurred so that the mouth movements would not hinder the evaluators' perception during the study.

Figure 6.4 User perception test video clip example of an interaction between a *SIA* (left) and a human participant (right).

Four videos (the agent displaying the behavior of the agent in one of the four conditions) are created for each of the 20 human-human video clips of the NoXi database. So, we have a total of 80 videos where the *SIA* replaces one of the human interlocutors (see Figure 8.2). The behaviors of the GT condition are also shown by replacing the selected human with the *SIA*. We use the same setting when comparing videos of the GT with videos of the computational models. As such we eliminate any impact a participant may have toward the virtual character Shiban et al. [2015].

Not to make an evaluation that lasts too long which may deteriorate the concentration of the perception study participants and thus hinder the study, we split the perception test into four groups. Each group has 5 human-human interaction video clips to evaluate (i.e. each participant evaluates 20 short videos of $7s$ of human-agent interaction for all four conditions). All the videos are shuffled so that their order does not impact our perception study.

For each perception test group, we recruit 30 participants and ask them to evaluate each video (20 videos per group) with the aforementioned set of questions. To filter out inattentive participants, for each video we randomly include attention check questions (e.g. "Is the virtual character playing tennis with the human interlocutor?").

Subjective Evaluation Results and Discussions

The participants' responses are grouped together according to their corresponding construct (behavior naturalness, behavior human-likeliness, synchrony, and engagement) for each condition (GT, our two baseline models of IL-LSTM and sym-IL-LSTM, and *ASAP*). We visualize the distribution for each construct, in Figures 8.3 and 8.4, and report the median values in Table 6.4.

One-way ANOVA reports significant differences among all animation conditions for all four constructs: behavior naturalness ($F = 41.5, p < 0.001$), behavior human-likeliness ($F = 43.1, p < 0.001$), synchrony ($F = 66.9, p < 0.001$), and engagement ($F = 90.0, p < 0.001$). A post-hoc pairwise comparison analysis is performed by running Tukey's honestly significantly differenced (HSD) test. Tukey's HSD reveals the following. Statistical significant differences were found between all pairs ($p < 0.001$) except for the pair of (sym-IL-LSTM, *ASAP*) for the constructs of behavior naturalness and human-likeliness ($p = 0.9$ and $p = 0.9$ respectively). Concerning the constructs of synchrony and engagement, all pairs were reported to be significantly different ($p < 0.003$). A two-tailed t-test was performed between all possible pairs of compared animations for each construct to test the statistical significance. The t-test p-values reported significant differences between all pairs ($p < 0.001$) except for the pair of (sym-IL-LSTM, *ASAP*) for the constructs of behavior naturalness and human-likeliness ($p = 0.7$ and $p = 0.5$ respectively). T-test yields significant differences among all conditions for synchrony and engagement constructs.

Methods	Naturalness	Human-likeliness	Synchrony	Engagement
GT	3.00/3.03	3.00/3.02	3.40/3.30	4.00/3.60
IL-LSTM	2.33/2.42	2.33/2.38	2.60/2.59	2.67/2.63
sym-IL-LSTM	2.67/2.63	2.67/2.59	2.80/2.78	3.00/3.01
ASAP (ours)	**2.67/2.66**	**2.67/2.63**	**3.00/2.97**	**3.33/3.24**

Table 6.4 Median/mean values of naturalness, human-likeliness, synchrony, and engagement.

From the subjective results, the simulation with GT values (median/mean) receives the highest values for all four constructs, namely behavior naturalness

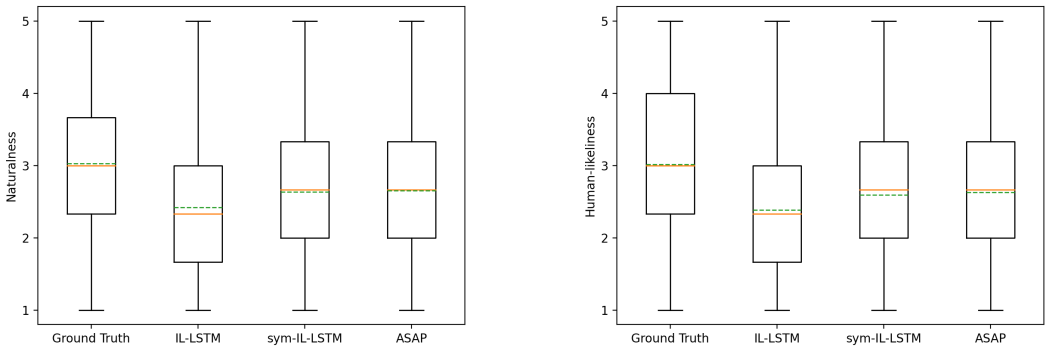


Figure 6.5 Distribution of behavior naturalness (left) and human-likeliness (right). Median represented by orange line and mean represented by green dashed line.

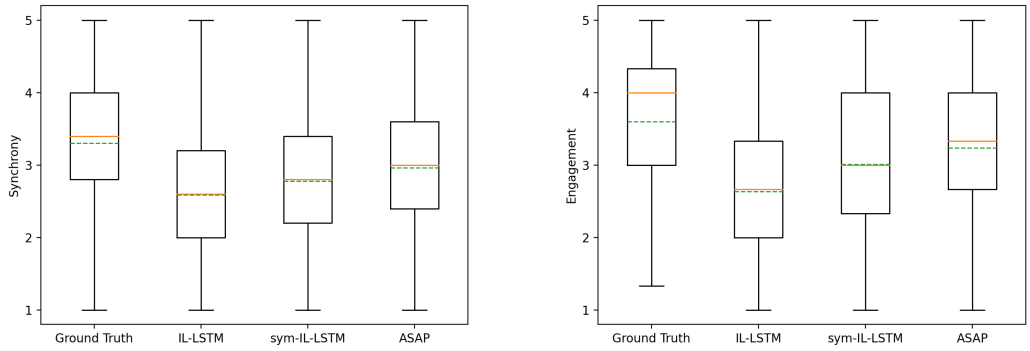


Figure 6.6 Distribution of synchrony (left) and engagement (right). Median represented by orange line and mean represented by green dashed line.

(3.00/3.03), behavior human-likeliness (3.00/3.02), synchrony (3.40/3.30), and engagement (4.00/3.60). Via the constructs of behavior naturalness and human-likeliness, a rise in quality can be noticed between that of IL-LSTM (2.33/2.42, 2.33/2.38 respectively) and the other two computational models of sym-IL-LSTM (2.67/2.63, 2.67/2.59 respectively) and our *ASAP* model (2.67/2.66, 2.67/2.63 respectively). We assume that this difference is due to the application of adaptive online prediction, instead of sliding window prediction as in IL-LSTM, which enables the generation of continuous motions which may lead to a higher perception of naturalness and human-likeliness. The quality of the generated agent behavior along the constructs of synchrony and engagement increases from the IL-LSTM (2.60/2.59, 2.67/2.63 respectively), to sym-IL-LSTM (2.80/2.78, 3.00/3.01 respectively), to *ASAP* (3.00/2.97, 3.33/3.24 respectively). We can remark that modeling of *reciprocal adaptation* allows *SIA* to be more in sync and engaged with its interlocutor.

We also want to evaluate if our *ASAP* model can produce behaviors for *SIA* being both a *listener* and a *speaker*. We check the quality of the generated agent

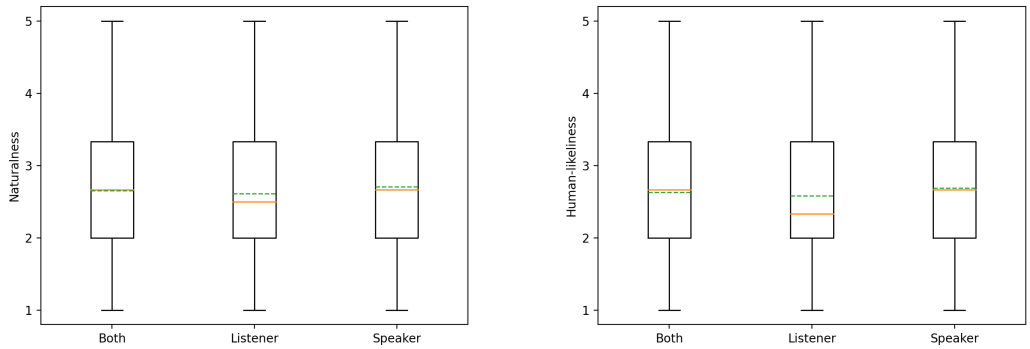


Figure 6.7 Distribution of behavior naturalness (left) and human-likeliness (right). Median represented by orange line and mean represented by green dashed line.

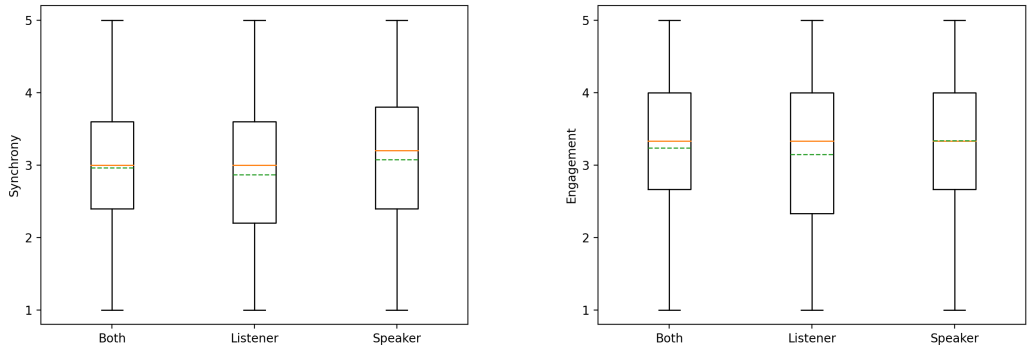


Figure 6.8 Distribution of synchrony (left) and engagement (right). Median represented by orange line and mean represented by green dashed line.

behavior of *ASAP* along the four constructs by comparing the produced behaviors as a *listener* and a *speaker*, as shown in Figures 6.7 and 6.8.

For the *SIA* being either a *listener* or a *speaker* or both combined, one-way ANOVA reported significant differences for the construct of synchrony ($p = 0.02$) but no significance for the other three constructs of behavior naturalness, behavior human-likeliness, and engagement. Tukey's HSD on synchrony revealed a significant difference between *listener* and *speaker* ($p = 0.01$). A two-tailed t-test was performed and showed significant differences between *listener* and *speaker* for the constructs of synchrony ($p = 0.005$) and engagement ($p = 0.02$).

We can remark that *ASAP* generates both *listener* (2.5, 2.3, 3.0, 3.3 respectively) and *speaker* (2.7, 2.7, 3.2, 3.3 respectively) behaviors with similar qualities which indicates that *ASAP* can be used to generate *SIA* behaviors for an entire interaction.

Our subjective evaluation results are inline with the results of the objective evaluation. Our *ASAP* model performs better than that of the baseline models of IL-LSTM and sym-IL-LSTM. Thus, *ASAP* outmatches the baselines along the four constructs (naturalness, human-likeliness, synchrony, and engagement), notably in terms of *synchrony* and *engagement*, and is the most similar to the GT both

quantitatively and qualitatively. Moreover, *ASAP* can serve to produce *SIA* behavior for both *speaker* and *listener*.

6.5 Contributions and Conclusion

6.5.1 Contributions

Our work makes the following contributions:

- We propose the modeling of *reciprocal adaptation* and show how the endowment of such capability can make *SIA*s behave more *social* and *engaged* as both *speaker* and *listener*;
- Our results show that *ASAP* out-performs state-of-the-art models quantitatively and qualitatively notably for interaction *synchrony* and *engagement*.

6.5.2 Conclusion

Having the goal to create an expressive *SIA* capable of interacting with the user while maintaining his/her attention, we develop a predictive model that produces the agent's nonverbal behaviors serving as both active *speaker* and *listener*. We modelize the *reciprocal adaptation* of our *ASAP* model by focusing on the aspects of *interpersonal temporality*, *multimodality* by encoding multimodal signals, and behavior prediction *continuity* with the autoregressive adaptive online prediction. Our model outperforms the baseline models through both objective and subjective evaluations. *ASAP* shows great promise in rendering *natural* and *human-like* behaviors that are *engaging* and *in sync* with the interlocutor addressing our two research questions.

The key points of this Chapter:

Addressing Research Questions

- The endowment of *reciprocal adaptation* capability improves *interpersonal dynamics* (synchrony and engagement) of the generated agent behaviors.
- The *quality* of the agent's behavior (naturalness and human-likeliness) is also enhanced with the modeling of *reciprocal adaptation*.

ASAP Model

- We modelized the *reciprocal adaptation* via *ASAP* model by focusing on the aspects of *interpersonal temporality*, *multimodality* by encoding multimodal signals, and behavior prediction *continuity* with the autoregressive adaptive online prediction.
- *ASAP* shows great promise in rendering *natural* and *human-like* behaviors that are *engaging* and *in sync* with the interlocutor.

Publication

- Jieyeon Woo, Catherine Pelachaud, and Catherine Achard. Asap: Endowing adaptation capability to agent in human-agent interaction. In *28th International Conference on Intelligent User Interfaces*, 2023d

Part IV

Real-time adaptive SIA system

Chapter 7

Adaptive SIA system for real-time human-agent interaction

Contents

In this Chapter, the design of an adaptive SIA system that provides *real-time* human-agent interaction is presented. The usefulness of the system is validated by applying it to a medical care application of Cognitive Behavior Therapy (CBT) as a proof-of-concept. The display of adaptive SIA behavior is shown to increase the user experience (user's impression of the agent) and the effectiveness of the chosen application.

7.1 Introduction

Adaptation is a key aspect of interpersonal relationships (Cappella [1991]). It can serve to indicate our *engagement* and *rapport* which can also elicit enhancement of involvement of others (Delaherche et al. [2012], Oertel et al. [2020], Gupta et al. [2019], Huang et al. [2010], Raffard et al. [2018]). Interlocutors adapt their behaviors throughout the interaction continuously, reciprocally, and dynamically to those of the others, referred to as *reciprocal adaptation* (ref. Chapter 2). The *reciprocal adaptation* arises between interlocutors in *real time* following a looped process.

Strengthening interpersonal relationships is important in any task in which people work together. In particular, in psychotherapy, including cognitive behavior therapy (CBT; Beck [2020]), the development of rapport (VandenBos [2007]) and the collaborative relationship between the supporter and the help-seeker has been emphasized, and their impact on treatment effectiveness has been investigated. Relationships of mutual understanding, acceptance, and sympathetic compatibility between or among individuals have been shown to contribute to the effectiveness of psychotherapy (DeVault et al. [2014], Huang et al. [2010], Raffard et al. [2018]). Adaptation, both verbal and nonverbal, strengthens the relationship between supporter and help-seeker, resulting in help-seekers feeling more at ease and more likely to confront their problematic relationships and improving adherence and persistence rate with the supporter's suggestions. In medical and psychological fields, it is known that rapport affects the effectiveness of CBT (Asay and Lambert [1999], Ardito and Rabellino [2011], Norcross and Lambert [2018]). During real therapy between a patient and a therapist, health support is provided through face-to-face interactions. The therapist not only provides the therapy through verbal communication but also expresses the feeling of sympathy and engagement non-verbally with their patient (Ramseyer and Tschacher [2014], Koole and Tschacher [2016]).

Virtual agents interact with human users by playing the role of interlocutor. Their central objective is to improve the human users' interaction experience by increasing their users' engagement level. A way to attain their goal is to adapt their behaviors depending on those of their users. For such embodied agents, they need to display *continuous* and *adaptive* behaviors in *real time*. Adaptive agents, adapting their verbal and/or nonverbal behavior, have demonstrated their use in increasing the user engagement (Schroder et al. [2011], Ritschel et al. [2017], Weber et al. [2018], Biancardi et al. [2021]), rapport (Huang et al. [2010], Raffard et al. [2018]), interaction synchrony (Raffard et al. [2018]), and impression of the agent (liking, naturalness, and human-likeliness) (Bailenson and Yee [2005], Huang et al. [2010], Biancardi et al. [2021]). The *real-time* aspect of behavior generation along with the fluid dialogue management needs to be assured throughout the whole interaction for both interlocutor roles of a listener and a speaker which is not a trivial task.

The use of virtual agents can be seen in various domains ranging from assistance (Sidner et al. [2018], Biancardi et al. [2021]) to healthcare (Philip et al. [2020], Bickmore [2022], Shidara et al. [2022]). Virtual agents have been demonstrated to be promising tools, notably for medical care, in gaining users' trust and

acceptance (Philip et al. [2020]). Several studies have highlighted the benefits of using virtual agents in e-health applications (Philip et al. [2020], Lisetti et al. [2013], Lucas et al. [2014], Bickmore [2022]). As such several conversational agents have been developed to deliver CBT focusing on the treatment (Ring et al. [2016], Fitzpatrick et al. [2017], Kimani et al. [2019], Shidara et al. [2022]). However, human supporters communicate with their help-seekers through behavior (both verbal and nonverbal) that are adapted to that of their help-seekers (Ramseyer and Tschacher [2008]). It is thus important for virtual CBT agents to also communicate verbally and non-verbally, and adapt their behavior to that of their users. It is not clear from previous studies whether behaviors with reciprocal adaptation, more specifically facial expressions and head movements, have an impact on improving the effectiveness of human-agent interaction during CBT.

The contributions of this study are twofold: (1) to elucidate whether adaptable virtual agents can enhance the *experience* perceived by users themselves, and (2) to ascertain whether adaptable virtual agents can improve the *effectiveness* of CBT through comparative experiments. We develop an adaptive virtual agent that renders adaptive behavior in *real-time* based on the behavior shown by human interlocutors. To generate the agent's behavior in our system, we adopted the *ASAP* model which renders mutually adaptive agent behavior (ref. Chapter 6). Our system loops through the processes of social signal perception, agent adaptive behavior generation, agent visualization, and signal transmission, ensuring *real-time* responsiveness. Furthermore, we demonstrate that non-verbal adaptation of virtual agents contributes to the achievement of interaction objectives in pairs of users and virtual agents. For this proof-of-concept, we incorporate a scenario based on CBT (Beck [1979, 2020]) into the virtual agent. CBT is an established mental healthcare method that provides face-to-face dialogues with users, similar to our system's setup. As CBT is effective not only for mental illnesses such as depression and anxiety disorders but also for coping with daily psychological distress, it becomes a suitable option for the general public (Greenberger and Padesky [2015]). In this study, we target the general population and conduct interactions based on CBT to cope with daily psychological distress. This research reveals how the adaptation of non-verbal behavior affects the relationship that users feel with the agent and how it impacts the objectives of the interaction. Additionally, we analyze how the relationship with the agent and the goals of the interaction are interrelated.

7.2 Related Works and Limitations

This section outlines relevant research on how the *reciprocal adaptation* of virtual agents has been modeled and can influence *user perceptions* in communication and the impact of *reciprocal adaptation* on mental health care, which is key to our interest.

7.2.1 Models of adaptation in HAI

Adapting to interlocutors is an essential part of interaction. Agents that interact with human users by taking the role of an interlocutor should also have the skill of adaptation. Related research has worked on creating conversational agents (virtual agents, humanoid robots, chatbots) that can adapt to their users. The adaptation can be done at different levels shown via various social cues (verbal, nonverbal, and/or conversational strategy) which are employed for diverse applications such as providing information, company, assistance, education, and medical care.

For the interaction to be personalized based on the interlocutor, several works have focused on adapting the agent's verbal context and/or conversational strategy. Nishimura et al. [2007] created a dialog system for chatbots that generates natural responses along with the response timing based on the user's inputs via decision trees. Ritschel et al. [2017] looked at the variation of linguistic style and its impact. The linguistic style was used to represent the robot's personality. The adaptation of linguistic style was modeled by a reinforcement learning model based on the user's engagement level which was estimated from the user's gaze and posture. Their study showed that adapting linguistic style can improve the user's engagement. Weber et al. [2018] studied the adaptation of jokes based on the user's sense of humor. The user's humor was detected without explicit user feedback through the user's smile and laughter. They proposed a robot that performs *real-time* adaptation using reinforcement learning and demonstrated that their robot performs significantly better in terms of amusement level by making jokes that consider its user's sense of humor compared to those that produce jokes randomly. Ding et al. [2022] created a conversational agent, TalkTive, that generates backchannels aiming to help the elderly to be engaged during cognitive assessments. TalkTive predicts the verbal backchanneling form that can be either reactive backchannels (e.g. "hmm") or proactive backchannels (e.g. "please keep going") and its timing.

The nonverbal channel plays a major role in communication. Other works have concentrated on modeling the nonverbal adaptation of agents. Huang et al. [2010] designed a virtual agent that produces visual backchannels via conditional random fields (CRFs) from the user's gaze, prosody, and lexical features. Their study denoted that visually adapted backchannels can reinforce the rapport agents build with their users and can be perceived as more natural. Schroder et al. [2011] also developed a virtual agent that produces nonverbal backchannels (smile, head movement, and vocalization). Agent displaying backchannels was able to engage its user better than an agent that does not. Agent's adaptation has also been expressed through the production of mimicry behavior. Bailenson and Yee [2005] created a virtual agent that renders the mimicry behavior. It imitates the user's head movements within a delay of $4sec$. A mimicking agent was shown to be perceived as more positive and persuasive than one that does not mimic the user. Raffard et al. [2018] also assessed the effect of mimicking virtual agents, with a mimicry delay varying between $0.5sec$ and $4sec$, with participants suffering from schizophrenia and healthy participants. Their results revealed that the rapport and interaction synchrony was improved with the display of mimicry behavior for

both participant groups. Adaptive nonverbal behavior is not only the one that is produced as a reaction to the user or copying the user's behavior but also the one that matches the user's behavior. Anderson et al. [2013] made a virtual agent framework for job interviews facilitating self-reflection and providing personalized coaching. The agent serves as a virtual recruiter that generates its nonverbal behaviors according to the user's face and hand gestures. Pecune et al. [2016] created a virtual agent for virtual tutor-child interaction that determines the agent's social goal (actions and communicative intentions) depending on its social role and its social relation toward the interacting child. Their study showed that the agent's role and social relation influence the agent's perception in terms of social attitude. Jones and Castellano [2018] proposed an adaptive robotic tutor for primary school children. Their robot based on open learner model (OLM) helped children to develop self-regulated learning (SRL) skills. Pereira Santos et al. [2023] built an embodied agent for obstetric simulation training. The agent playing the role of a digital patient adapts its facial expressions in *real time* via the behavior that is commanded by an on-screen controller. Sidner et al. [2018] realized a *real-time* architecture for companion agents (virtual agents and robots) that provides companionship for the elderly. The agent adapts its gesture via the user's facial gestures and motions. Biancardi et al. [2021] created a virtual agent that adapts its behaviors to that of its interlocutor. It serves as a virtual museum guide and aims to maximize the user's engagement. Their virtual agent is capable of adapting its nonverbal cues at the behavioral and conversational levels. They demonstrated that the adaptive agent was perceived as more positive than the non-adaptive agent.

7.2.2 Mental health care with virtual agents

CBT is an effective and well-established healthcare method for addressing mental illnesses such as depression and anxiety and for daily health care. Although CBT is effective, its lack of accessibility is a serious issue, as it requires a high level of skill for supporters. Various types of conversational agents have been employed to promote the use of CBT. Conversational agents include text-based agents, such as those used in messaging applications, robotic platforms, and virtual agents. Among them, text-based agents are particularly common due to their ease of use and are often used as smartphone applications for convenient access. Two examples of such applications are Wysa (Inkster et al. [2018]) and Woebot (Fitzpatrick et al. [2017]), which have been tested for their ability to assist individuals exhibiting mild to moderate depression and anxiety symptoms. These tools are designed to provide regular mental health care interventions rather than clinical treatment. On the other hand, one limitation of these text-based agents is that they communicate only via language. In mental health care, interactions using nonverbal behavioral modalities are important for improving the relationship between the supporter and the help-seeker, as they influence various aspects, such as understanding the psychological state of the other person, empathic behavior, and sense of presence.

Virtual agents and robots have the advantage of face-to-face multimodal interaction, including facial expressions, gestures, and voice. In particular, virtual

agents can be used at a relatively low cost, and their appearance and voice can be customized. For these reasons, virtual agents are anticipated for use in mental health care. DeVault et al. [2014] conducted experiments of medical interviews with virtual agents that operated using the Wizard-of-OZ method or autonomously. The results showed that users felt a rapport with the agent. It was also suggested that an autonomous agent without an operator behind them could facilitate user self-disclosure. In addition, attempts are being made to use virtual agents not only for medical interviews for screening purposes but also for mental health care, such as CBT. Shidara et al. [2022] developed a virtual agent that delivers CBT which helps users to identify and rectify automatic thoughts. They alternated the conversational strategy for participants who needed help in identifying their automatic thoughts via a language model-based automatic thought classifier. Ring et al. [2016] also illustrated a virtual therapist agent for CBT counseling. The system dialogue is managed with the user's speech and affect (detected from the user's speech). Along with the dialogue management, they generate the agent's nonverbal behavior, which is automated via the Behavior Expression Animation Toolkit (BEAT; Cassell et al. [2001]). The potential efficacy of affect-aware agents for the guidance of CBT scenarios is presented. Efforts have also been made to improve user receptivity by adjusting the appearance of the agent and its set age, etc. Parmar et al. [2022] have systematically manipulated animation quality, speech quality, rendering style, and simulated empathy in the domain of health counseling. They investigate the effects on virtual agents' perceptions in terms of spontaneity, engagement, trust, credibility, and persuasion. The results showed that the agents improved their ability to persuade but hindered their ability to improve trust. In terms of agent design, suggestions include agents tailored to black church communities (O'Leary et al. [2020]) and agents with the appearance of older adults (Razavi et al. [2022]) to provide realistic conversational practice to older adults at risk for isolation and social anxiety. While adaptations such as these have been made, the mental health impact of *real-time* adjustment of nonverbal behaviors such as facial expressions and head movements is not clear.

In this study, we seek to develop a virtual agent that is capable of adapting its behavior to its interacting user. We focus mainly on the nonverbal adaptation of the agent generating expressive and adaptive agent's facial expressions and head-/gaze movements. We integrate the *ASAP* model (ref. Chapter 6) to enable the reciprocal adaptation to virtual agents. We intend to check how human interactants perceive adaptive agents in terms of naturalness, human-likeliness, interaction synchrony, and engagement. We aim to also adapt the conversational strategy (or conversational move) for our selected scenario, the CBT scenario, to the user's response. Compared to the previous works, we tried to assure the *real-time* functioning of our agent at the frame-level to display *continuous* agent movements. Moreover, to our knowledge, no non-verbal adaptive agent has yet been introduced for the application of CBT. We intend to improve the interaction itself, more precisely the user experience (perception of the agent), and at the same time the effect of CBT compared to previous CBT systems (Ring et al. [2016], Shidara et al. [2022]) by rendering adaptive agent behavior.

7.3 Method

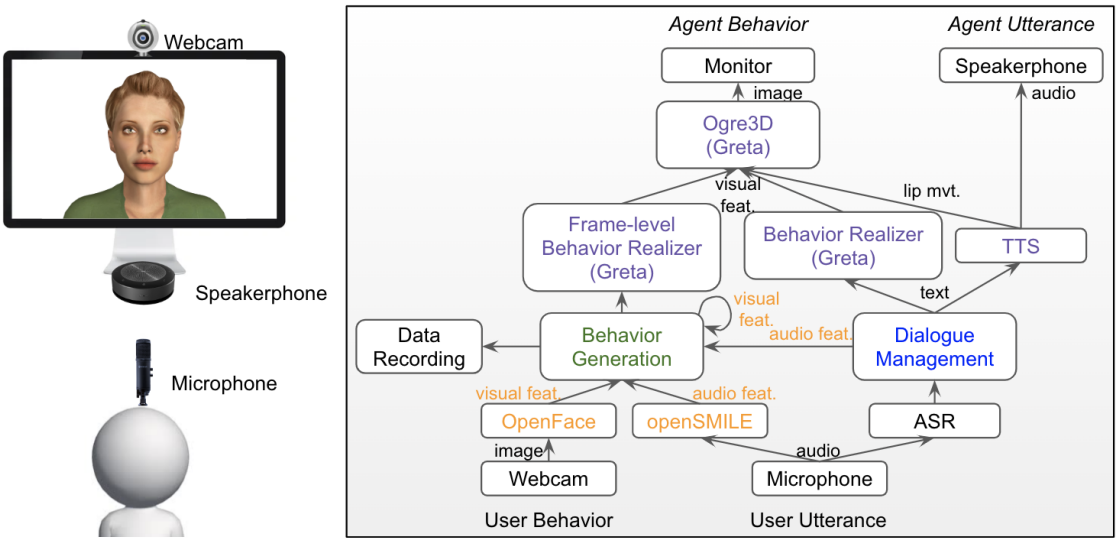


Figure 7.1 Real-time expressive and adaptive agent system setup (left) and architecture (right). The proposed system interacts with the user via a virtual agent that shows expressive and adaptive behavior in *real time*. It captures the user's face with a webcam and the user's speech with a microphone. The agent is displayed in front of the user on a monitor and its speech is rendered via a speech synthesizer and a speakerphone. Consists of 4 main functionalities: perception of the user's and agent's own behavior (in orange), generation of expressive and adaptive behavior (in green), dialog management (in blue), and visualization of the agent's behavior (in violet)

To improve the user's interaction experience and the perception of the agent, we built a *real-time* expressive and adaptive virtual system using the *ASAP* model, to render reciprocally adaptive agent behaviors. Our system, illustrated in Figure 7.1, is composed of 4 main functionalities: perception of the user's and agent's own behavior (via OpenFace, openSMILE, and ASR), generation of expressive and adaptive behavior (via behavior generation module), dialog management (via dialogue management module), and visualization of the agent's behavior (via frame-level behavior realizer and behavior realizer modules and Ogre3D). For the effectiveness assessment of our system, CBT is chosen as our proof-of-concept to validate our experimental hypotheses stated in the upcoming sections.

7.3.1 Real-time Expressive and Adaptive Agent System

For an agent to be capable of interacting in *real time* and displaying expressive and adaptive behavior, it needs to possess the following functionalities:

- perception of the user's behavior;
- perception of its own behavior;
- generation of expressive and adaptive behavior;

- dialog management;
- visualization of its behavior.

We build an expressive, interactive, and adaptive virtual agent system, *IAVA* system (see Appendix B in Chapter 11 for more detail), that provides *real-time* human-agent interaction. Each of the functionalities is endowed to our agent by employing the corresponding techniques. We present in detail each functionality.

Perception of User's Behavior

An agent interacting with a human user needs to act accordingly to its user. It needs to take into account the user's behavior (speech and gesture). To do so, the agent first needs to perceive such signals. The perception can be separated into three parts:

- User's speech content;
- User's speech prosody;
- User's gesture.

User's speech content The user's speech content is important as the user's intentions are explicitly expressed via words. This information is essential for all types of automatic systems interacting with the user which is easily seen in conversational AI assistants such as Google, Alexa, and Siri. To capture this information of speech content, we choose to integrate the Google ASR [1]. Automatic Speech Recognition (ASR), also referred to as Speech-to-Text (STT), is the transcription technology that captures the audio of spoken words and transforms it into written text. The ASR technique identifies full phrases and transcribes them as the user is speaking.

User's speech prosody When we speak, our intentions are conveyed via the speech content but we also vary the way we speak using our voice qualities. Our speech can vary in pitch (high or low), loudness (loud or soft), and duration (fast or slow). This variation in speech is referred to as speech prosody. We extract the speech prosody using openSMILE (Eyben et al. [2010]), an open-source toolkit for audio feature extraction. We choose to retrieve the prosodic features of:

- Fundamental frequency representing the pitch;
- Loudness quantifying the sound energy;
- Voicing probability estimating a percentage of unvoiced and voiced energy;
- Mel-frequency Cepstral Coefficient (MFCC; Logan [2000]) which is a representation of the short-term power spectrum of a sound.

These features are obtained in *real-time* with a frequency of $100Hz$.

[1] https://cloud.google.com/speech-to-text

User's gesture For the gesture, we decide to focus on the perception and the generation of the facial gestures (facial expressions and head/gaze movements) of both the human user and the agent. For the perception of the user's facial gestures, we use OpenFace (Baltrušaitis et al. [2016]), an open-source toolkit, to extract the facial features at the frame-rate of $30fps$ which are as follows.

- Gaze movements (G_x and G_y) which are the gaze angles w.r.t. the x and y axis;
- Head movements (R_x, R_y, and R_z) which are the Euler head rotations w.r.t. the x, y and z axis;
- Facial expressions via facial Action Units (AUs; Ekman and Friesen [1976]) which are facial muscle movements defined by the Facial Action Coding Systems (FACS; Ekman and Friesen [1978]). We use *AU1* (inner brow raiser), *AU2* (outer brow raiser), *AU4* (brow lowerer), *AU5* (upper lid raiser), *AU6* (cheek raiser), and *AU7* (lid tightener).

Perception of Agent's Behavior

The agent needs to behave not only depending on its user's behavior but also with respect to its previously displayed behavior and its current intention (e.g. agent utterance). For the agent to act as such, it needs to also perceive (or remember) its own behavior (i.e. the agent behavior at timestep t takes into account its previous behavior until timestep $t-1$). For this, the agent remembers its previous behaviors (speech content and prosody computed in *real time* via openSMILE at $100Hz$, and previously rendered facial gestures) within its internal memory along with its currently exhibiting behavior.

Generation of Expressive and Adaptive Agent's Behavior

The behavior signals that are perceived by the human user and the agent are used to generate expressive and adaptive agent behavior. We employ the Augmented Self-Attention Pruning (*ASAP*) model (ref. Chapter 6). This model generates expressive agent facial gestures that are reciprocally adaptive. It learns interpersonal relationships via real human-human interactions, from a corpus of screen-mediated face-to-face interactions, the NoXi database (Cafaro et al. [2017]), with its self-attention pruning and data augmentation techniques. It also assures the movement continuity of the generated behavior by performing autoregressive adaptive online prediction.

The pre-trained *ASAP* model is integrated into our system. The perceived visual and audio features of the past 100 time-steps of both human user and agent are passed to the *ASAP* model to render the agent's expressive and adaptive visual behavior at the next time-step. The communication protocols of ZeroMQ [2] (Hintjens [2013]) and OSC (Open Sound Control) [3] (Wright [2005]) are used to pass the signals (for visual and audio signals respectively), after syncing the different

[2] https://zeromq.org
[3] https://opensoundcontrol.org

sampling rates to $25Hz$ which is the *ASAP* model's sampling rate, to the model for the prediction. The model generates the agent's behavior with an inference speed of $0.008s$.

Dialog management

During a conversation, the dialog is managed between interlocutors by taking turns, and the next flow of discussion is decided depending on the speech content. The agent also needs to be able to behave similarly in a *real-time* conversation with its human user. To this aim, we use Flipper2.0 (or Flipper; van Waterschoot et al. [2018]) which is a dialog engine that can flexibly direct the conversational flow. The utterance text of the user's speech is obtained from the ASR and passed to Flipper via ActiveMQ [4](Snyder et al. [2011]), a communication protocol, to choose the next conversational move based on a rule-based structure.

Visualization of Agent

For the visualization of the animation of the virtual agent, we use the Greta platform (Niewiadomski et al. [2009]) which is an open-source virtual agent platform simulating an agent's verbal and nonverbal behavior in *real time*. The agent's speech (its next conversational move chosen by the Flipper dialog engine) is transformed from text to audio using the CereProc [5] speech synthesizer (or Text-to-Speech (TTS)) within the internal audio module. The matching mouth movements are also produced along with the speech audio by the internal behavior realizer module. For the display of the agent's adaptive facial gesture, generated by the integrated *ASAP* model, the predictions are passed to the frame-level behavior realizer module. The agent's behavior outputted by the behavior realizer modules (adaptive agent behavior and mouth movements) is then passed to Ogre3D [6], an open-source scene-oriented 3D rendering engine, for display.

System Setup and Performance

For the system setup, a virtual agent is displayed in front of the user on a monitor (in a close-up of their face, head, and shoulders), as depicted in Figure 7.1. The user's speech is captured via a microphone and the user's facial gestures (head and gaze movements, and facial expressions) are obtained through a 1080p RGB webcam. The agent's spoken utterance (speech of the chosen conversational move) is rendered using a speech synthesizer and a speakerphone.

The system runs on two computers in parallel. The first computer continuously displays the agent's behavior via the Greta platform. The second computer runs the *ASAP* model, generating expressive and adaptive agent behavior in *real time*, along with the perception toolkits of OpenFace and openSMILE (for facial feature extraction and prosodic feature extraction respectively).

[4]https://activemq.apache.org
[5]https://www.cereproc.com
[6]https://www.ogre3d.org/

For the system's performance, *real-time* functioning ($25Hz$) is assured. A single system loop execution time of $0.04s$ with no delay is assured. To detail, the system loop consists of perception (approx. $0.03s$), behavior generation (approx. $0.008s$), communication (approx. $0.001s$), and visualization (approx. $0.001s$) which are all synced.

To run the system, there is a space requirement of approximately $7GB$ for the setup and use which consists of $2GB$ for platform visualization, $2GB$ for OpenFace and openSMILE, and $3GB$ for execution and data saving. Hardware specifications are: 2 computers with $2.4GHz$ Intel Core i9 mounted with NVIDIA Quadro RTX 4000 and $64GB$ RAM.

7.3.2 Cognitive Behavior Therapy

To assess if endowing virtual agents with reciprocal adaptation mechanisms enhances their user's experience, we choose to use the CBT scenario as our proof-of-concept.

CBT (Beck [1979, 2020]) is a mental health treatment that restructures automatic thoughts (or irrational thoughts). These automatic thoughts are those that come up to our minds suddenly and unconsciously. Because of their nature of occurring unexpectedly, we are not aware of them but they affect our mood. These thoughts often elicit negative feelings but can also evoke misleading positive emotions. To help people to recognize and rectify automatic thoughts into balanced ones, the CBT treatment is delivered. The restructuring of thoughts is done by asking the participants several fact-finding questions to guide them through the process of identification of such thoughts and changing them. The key effect of CBT is that it clarifies irrational situations and thoughts, and brightens people's moods.

By applying our adaptive virtual agent to CBT, we expect to see an amelioration in the effectiveness of CBT in mood improvement along with the user's experience, notably the enhancement of the agent's perception.

Scenario

We choose to work with the CBT scenario presented in Shidara et al. [2022] which is presented in Figure 7.2. We follow the same scenario (translated to French) of asking the participants to self-report their mood (negative mood intensities asked before and after the CBT session via questions Q3 and Q14), helping them to identify their automatic thought (via question Q4), and guiding them with fact-finding questions (by asking questions concerning proof, disproof, disputing, Socratic questioning, balanced thought, and caring words). For the identification of participants' automatic thought, their response is verified by an automatic thought classifier model (detailed below) to check if it corresponds to an automatic thought. The model performs a classification on the participants' response by checking whether it is based on negatively distorted cognition or factual validity to judge if they have correctly identified an automatic thought. The next conversational move (agent's utterance) is selected depending on whether the user's answer is an automatic thought or not.

No.	Item	Sentence (Virtual agent's questions)
		Hello, my name is Mai, and I'm a therapist. I'm here to help you learn how to face your worries.
Q1	Situation	*Has anything been bothering you lately, maybe something that's difficult to deal with or to face?*
		Please describe anything that is painful or burdensome.
Q2	Mood	*How did your mood change that time?*
Q3	Mood score (0–100%)	*How would you describe your mood's intensity from 0 to 100?*
Q4	Identification of an automatic thought	*I want to know why you felt like that in that situation.*
		What thoughts came to you when you faced that event?
		I see...
Q5	Proof	*Such thoughts that surface when we face certain events are called automatic thoughts.*
		Since automatic thoughts are unconscious, they sometimes seem true.
		If your automatic thought is accurate, what do you think it is based on?
Q6	Disproof	*So, do you have any other thoughts about that situation?*
		For example, does anything contradict your automatic thoughts?
Q7	Disputing(confirmation)	*Do you have any other thoughts? If so, can you share them?*
Q8	Socratic questioning* 1	*Let's examine the basis of the automatic thought that we have considered so far as well as how thoughts are balanced by contrary facts. Let me ask you some questions.*
		What is the worst possible consequence of this situation?
Q9	Socratic questioning 2	*Looking at the other side, what is the end result when you act as you want?*
Q10	Socratic questioning 3	*Please describe the most realistic scenario based on the predictions you just made.*
Q11	Balanced thought	*From that answer, let's create a new thought. How can you rethink this event?*
Q12	Caring words 1	*Good. Do you have anyone who will listen to you talk about this problem?*
Q13	Caring words 2	*I see, can you manage that?*
Q14	Mood score after change (0 to 100%)	*Well, our time is up for today. Thank you for your hard work. Please feel free to call again.*
		Please feel free to talk to me. How has the intensity of your original mood changed?
		Express this on a scale from 0 to 100.
[End]	Concluding a session	*If it's different from the beginning, that suggests that you internalized my suggestions well.*
		That's it for today. Thank you for your hard work. Please feel free to call again.

*Questions that create awareness of inconsistency in thoughts.

Figure 7.2 Scenario of utterances spoken by system used in Shidara et al. [2022] under the copyright terms of CC BY.

Automatic Thought Classifier

Training data	Test data	Feature	Accuracy	F1-score
D	D	TF-IDF	0.73	0.81
		BERT	0.67	0.76
		TF-IDF + BERT	0.67	0.76
G	D	TF-IDF	0.70	0.79
		BERT	0.70	0.79
		TF-IDF + BERT	0.67	0.78
D + G	D	TF-IDF	**0.76**	**0.83**
		BERT	0.73	0.79
		TF-IDF + BERT	0.73	0.79

Table 7.1 Classification results of automatic thought for French. D: data collected in Shidara et al. [2022], G: sentences from Greenberger and Padesky [2015], and the score in bold represents the best score.

The automatic thought classifier is a classification model that serves to validate whether the participants have successfully identified their automatic thought. The model is based on the Support Vector Machine (SVM) classifier algorithm with

a linear kernel (Shidara et al. [2022]). We get the French word embeddings from a pre-trained language model for word representations, the Bidirectional Encoder Representations from Transformers (BERT; Devlin et al. [2018]), and use the embeddings to train the automatic thought classifier. The choice of using BERT embeddings was chosen after testing different combinations of training data and distributed representations of Term Frequency-Inverse Document Frequency (TF-IDF), BERT, and both TF-IDF and BERT. The classification results for French are reported in Table 7.1 which shows that TF-IDF has the best accuracy and F1-score of 0.76 and 0.83 respectively when evaluated on sentences from Greenberger and Padesky [2015] (e.g. "If I'm not a total success, I'm a failure." and "I'll be so upset, I won't be able to function at all.") and trained with both data collected in Shidara et al. [2022] and sentences from Greenberger and Padesky [2015]. However, through supplementary evaluation of French automatic thought sentences (more complex and longer versions of sentences from Greenberger and Padesky [2015]), the classification with TF-IDF failed (none of the sentences in the supplementary evaluation was correctly classified) while the one with BERT performed with similar accuracy and F1-score of 0.73 and 0.79 respectively. Thus, BERT was selected as our input feature for the automatic thought classifier.

When inputting the participant's response (user's utterance), the model identifies automatic thoughts by performing binary classification. If the response was identified by the classifier as an automatic thought, the agent moves on to the next item of the scenario. However, if it was determined not to be an automatic thought, the agent provides a hint and asks the user again for an automatic thought. The agent provides at most six hints (from Beck [2020]) and if the user was unsuccessful to answer with an automatic thought even after six attempts, failing for the seventh time, the agent moves on to the next item. The automatic thought classifier is integrated within the dialogue management module (ref. Section 7.3.1 and Figure 7.1).

7.3.3 Experiment

Hypotheses

To investigate the usefulness of a virtual agent showing expressive and adaptive behavior, we propose three experimental conditions which are as follows:

- RA: Reciprocal Adaptation;
- MM: Mismatched Movement;
- SP: Still Posture.

Each condition is detailed in Table 7.2.

We build our study upon the work of Shidara et al. [2022] which demonstrated that interacting with virtual agents helps to improve the effectiveness of CBT (in improving people's mood and state).

For our study, we hypothesize that adaptive and expressive agents can further enhance the user experience (agent perception in terms of behavior naturalness, human-likeness, synchrony, and engagement) and the effectiveness of CBT (user

Experimental Conditions	Characteristic
Reciprocal Adaptation (RA)	Agent showing **expressive** and **adaptive** behavior; w/ ASAP model.
Mismatched Movement (MM)	Agent displaying pre-registered mismatched behavior; **expressive** but **not adaptive**.
Still Posture (SP)	**Still** posture with only lip-sync; w/o ASAP model Shidara et al. [2022].

Table 7.2 Experimental conditions.

mood and state change) based on the previous observation that virtual agents with adaptation help improve the interaction. Thus, we expect agents enabled with reciprocal adaptation (RA) to outperform agents showing random behaviors (or mismatched behavior; MM) and still agents (or in a still position; SP).

Our hypotheses are the following:

- H1: Reciprocally adaptive agents are more positively perceived and more effective for CBT compared to agents showing mismatched behaviors (RA $>$ MM);
- H2: Reciprocally adaptive agents are more positively perceived and more effective for CBT compared to agents in a still position (RA $>$ SP).

Protocol of Experiment

For the experiment of the interaction of human users with our agent system to proceed smoothly, we conduct our experiment with the following protocol:

i) Oral explication.

ii) Signing consent forms.

iii) Lecture of a document on automatic thought.

iv) Filling out pre-questionnaires.

v) Interaction between the participant and our virtual agent.

vi) Filling out post-questionnaires.

The detail of the protocol is as follows. We start the experiment by giving the participants an oral explication of the aim of the study and the experimental protocol (pre-questionnaire, interaction with our agent, and post-questionnaires) is given to the participant. The participants sign two consent forms: one to participate in the experiment and the other to give the authorization of use concerning the collected data. The participant is then invited to read a document explaining what an automatic thought is for a better understanding of the CBT scenario. They start by filling out the pre-questionnaires before the interaction. After the interaction with our virtual agent, they finish by filling out post-questionnaires.

Evaluating Questionnaires and Measures

To assess the potency of our system especially for our proof-of-concept of the CBT scenario and validate our hypotheses, we use questionnaires and objective measures from the literature.

Questionnaires For the evaluation of the perception of the agent, on a 5-point Likert scale (from 1 (not at all) to 5 (very)), we formulate our questionnaires based on existing questionnaires of human-agent interaction evaluation which are the following:

- Behavior naturalness (Fitrianie et al. [2021], Von der Pütten et al. [2010]): e.g. "Is the behavior of the virtual agent artificial?";
- Behavior human-likeliness (Fitrianie et al. [2021], Von der Pütten et al. [2010]): e.g. "Does the virtual agent behave like a human?";
- Engagement (Fitrianie et al. [2021], Von der Pütten et al. [2010]): e.g. "The virtual agent was engaged in the conversation?";
- Synchrony (Prepin et al. [2013], Louwerse et al. [2012]): e.g. "The virtual agent and I were agreeing to each other?";
- Rapport (Wang and Gratch [2009], Von der Pütten et al. [2010]): e.g. "I think the virtual agent and I established a rapport.".

The same questionnaires were used for the agent perception constructs of behavior naturalness, behavior human-likeliness, engagement, and synchrony as in Chapter 6 (questions at the third person point of view; e.g. "Are the human and the virtual character/agent agreeing to each other?") with questions reformulated at the first person point of view (e.g. "The virtual agent and I were agreeing to each other?").

As CBT is a psychological therapy, to assess the effectiveness of CBT we take the questionnaires from the psychology field which are as follows.

- State-Trait Anxiety Inventory (STAI; Spielberger et al. [1971]): psychological inventory consisting of 40 self-report items for measuring participant's anxiety level (state and trait with 20 items each scored from 1 (not anxious at all) to 4 (very anxious)), e.g. "I feel nervous.";
- Kessler Psychological Distress Scale (K6; Kessler et al. [2002]): six-item self-report measure via a 5-point Likert scale (ranging from 0 (none of the time) to 4 (all of the time)) for measuring participant's psychological distress level, e.g. "How often did you feel so depressed that nothing could cheer you up?";
- Cognitive Change-Immediate Scale (CC; Schmidt et al. [2019], Vittorio et al. [2022]): five-item self-report measure (rated on a scale from 0 (not at all) to 6 (completely)) that assesses help-seekers' experience of cognitive change and cognitive skill use during sessions, e.g. "I noticed myself thinking less negatively.".

The anxiety state change is calculated via STAI-state score (total score of 20 STAI state items), which is reported before and after the CBT session, using Equation 7.1.

$$Anxiety\,state\,change = STAI\text{-}state_{pre} - STAI\text{-}state_{post} \tag{7.1}$$

For psychological distress, the participants reported their distress level using the K6 scale before and after the session. Before the experiment, the participants responded to the K6 scale taking into account how they felt over the previous 30 days. After the experiment, they were asked to report whether they felt a change for any of the K6 items after the CBT session. The change in psychological distress level is measured via K6 score (total score of 6 K6 items) by Equation 7.2.

$$Psychological\,distress\,level\,change = K6_{pre} - K6_{post} \tag{7.2}$$

We also measure the mood change by computing the mood change score, defined in Equation 7.3, presented in Shidara et al. [2022]. The mood scores ($mood_{pre}$ and $mood_{post}$) are self-reported mood scores (negative mood intensities between 0-100; 0 for happy and 100 for depressed).

$$Mood\,change = \frac{(mood_{pre} - mood_{post})}{mood_{pre}} \tag{7.3}$$

Measures To fully evaluate the performance of our system, we also assess it via objective measures. As we focus on the adaptation, it is interesting to investigate whether it was established within the interaction between our agent and the user along with the evaluation of behavior appropriateness. We employ the measures of Kolmogorov-Smirnov two-sample test (KS test) and DTW, presented in Chapter 6, along with reciprocal adaptation measures of Synchrony (Sync) and Entrainment Loop (EL), introduced in Chapter 5, between the behaviors of the user and the agent under one of the experimental conditions (RA, MM, or SP).

For the DTW, we check the proximity/resemblance between the agent's generated behavior (for one of the experimental conditions of RA, MM, or SP) and the user's behavior against that of the human-human interaction (interactions in the NoXi database (Cafaro et al. [2017]); ref. Chapter 4) to evaluate reciprocal adaptation. To detail, to assess the CBT session, the behaviors of the user and the agent under one of the experimental conditions (RA, MM, or SP) are used to compute the proximity score. This CBT session score is compared with the human-human interaction score which is the average of the scores obtained from the pairs of interlocutors within NoXi.

We also computed the quantity of movement (ΔQ_{mvt}) which is the quantity of movement (head rotations R_x, R_y, and R_z) defined as in Equation 7.4:

$$\Delta Q_{mvt} = \sum_{t=1}^{T} \sqrt{(R_{x,t} - R_{x,t-1})^2 + (R_{y,t} - R_{y,t-1})^2 + (R_{z,t} - R_{z,t-1})^2} \tag{7.4}$$

where T is the sequence length.

Statistical Analysis A one-sided Welch's t-test was employed to compare the measuring aspects of user experience and CBT effectiveness for the experimental group pairs of (RA,MM) and (RA,SP) to examine our two hypotheses.

Interpretation The perception of the agent factors of behavior naturalness, human-likeliness, synchrony, engagement, and rapport are interpreted as rendering a more positive agent impression when the score is closer to 5.

For the interpretation of the agent's quantitative measure, the lower the values are for the metrics of Q_{mvt} and KS test, the better the agent performs as it is closer to the human interlocutor within a human-human interaction. Concerning the measures of synchrony and entrainment loop, the higher the value the better they are.

For the measures of CBT, the CBT is effective, showing the improvement of a certain factor, when:

- Mood: mood score (negative mood from pre to post) decreases or high mood change score;
- Anxiety: anxiety state (anxiety level from pre to post) decreases or high anxiety state change score;
- Psychological distress: psychological distress level (stress level from pre to post) decreases or high psychological distress level change score;
- Cognitive change: high CC score (experience of cognitive change and cognitive skill).

Experimental Setting

A sample size of 60 French-speaking participants (confidence level of 90 with a margin of error of 10 for the French population with a higher education degree (population portion of 33%)). We separate them into our 3 experimental conditions of RA, MM, and SP having 20 participants for each condition. Table 7.3 presents the demographics of the recruited participants per experimental condition (RA, MM, and SP). The inclusion criteria for this study were French-speaking, 18 years or older, and no vision or hearing impairments.

	RA	MM	SP
Number of participants	20	20	20
Gender			
Male	14	11	15
Female	6	9	5
Age			
Under 30	17	15	16
Above 30	3	5	4

Table 7.3 Participant demographics per condition (RA, MM, and SP).

7.4 Results

To validate our hypotheses, we study the impact of our *real-time* adaptive agent on the perception of the agent, effectiveness for CBT (user mood and state change), and the relation between the agent perception and user mood/state change.

7.4.1 Perception of Agent's Behavior

We also hold interest in whether our experimental conditions (RA, MM, and SP) influence how the participants perceive the agent (perception of the agent's behavior) along the 5 factors of naturalness, human-likeliness, synchrony, engagement, and rapport.

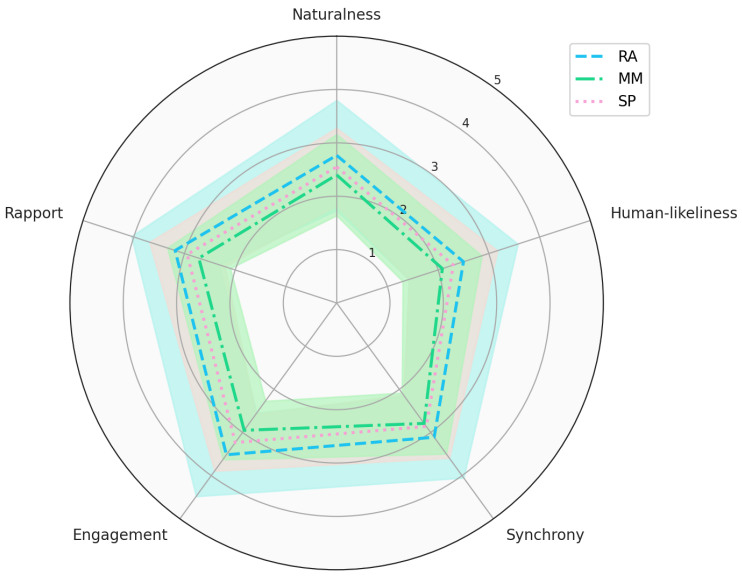


Figure 7.3 Perception of the agent's behavior along naturalness, human-likeliness, synchrony, engagement, and rapport. The central line in bold represents the mean value of each condition (RA, MM, and SP) and the colored-filled contour represents the standard deviation of each condition.

For the perception of the agent's behavior, we can note, in Figure 7.3 and Table 7.4, that the condition of showing reciprocally adaptive behavior (RA) is perceived as the most natural, human-like, in sync, engaged, and to have a rapport with the participant among the three experimental conditions. An agent showing reciprocally adaptive behavior is indeed perceived more positively along all 5 aspects compared to an agent displaying random behavior (or mismatched behavior; MM) validating our first hypothesis H1 (RA $>$ MM). We also validate that RA condition is better than a still agent (or in a still position; SP) also for all 5 aspects validating our second hypothesis H2 (RA $>$ SP). A significant difference was found between the experimental conditions of RA and MM for the engagement

Factor	RA		MM		SP	
	Mean	SD	Mean	SD	Mean	SD
Naturalness	2.77	1.030	2.40	0.762	2.55	0.728
Human-likeliness	2.50	1.070	2.08	0.779	2.30	0.871
Synchrony	3.11	0.941	2.79	0.706	2.86	0.732
Engagement	3.52	0.964	2.95	0.678	3.23	0.659
Rapport	3.18	0.860	2.72	0.620	2.95	0.743

Factor	Between RA-MM		Between RA-SP	
	Cohen's d	P-value	Cohen's d	P-value
Naturalness	1.28	0.210	0.767	0.448
Human-likeliness	1.41	0.169	0.647	0.522
Synchrony	1.22	0.232	0.938	0.355
Engagement	2.15	0.039	1.085	0.286
Rapport	1.97	0.057	0.935	0.356

Table 7.4 Mean and standard deviation of agent behavior perception factors for the conditions of RA, MM, and SP. Significance difference between the condition pairs (RA,MM) and (RA,SP) reported via one-sided Welch's t-test.

aspect (Cohen's d=2.15 and p-value=0.039). No significance was found for the other aspects and for the aspects between RA and SP conditions.

It is interesting to notice that the least performing condition for all 5 factors is the MM condition which shows random and unsynced agent behaviors. It seems that the display of random behavior disregarding the interlocutor rather hinders the impression of the agent. Maintaining a still posture (SP) is preferred to showing non-adaptive behaviors that not considering the participant at all.

We also look at objective measures of DTW, KS test, and synchrony and entrainment loop measures. For the quantitative study, as we want to see the influence of having adaptive behaviors, we compare our conditions of adaptive (RA) and non-adaptive (MM).

We can check in Figures 7.4 and 7.5 and Table 7.5 that the condition displaying reciprocally adaptive behavior (RA) performs better than that showing non-adaptive or mismatched behavior (MM) in terms of head movement quantity (via Q_{mvt}), density distribution similarity (via KS test), and DTW resemblance. In addition, we can remark that the adaptive condition (RA) is better than the non-adaptive condition (MM) in being in sync and in entraining its interlocutor's behavior. Via these quantitative measures, we can also note that an adaptive agent scores better than a non-adaptive one further supporting our prior observation that the RA condition renders a better impression than the MM condition.

7.4.2 User Mood and State Change

We investigate the global effect of CBT for all three experimental conditions (RA, MM, and SP) combined, with a total number of 60 participants, and the effect of each experimental condition. We assess the effect of the CBT experiment by looking at the factors of mood, anxiety, psychological distress, and cognitive change.

Measure	RA		MM		Cohen's d	P-value
	Mean	SD	Mean	SD		
ΔQ_{mvt}	12.4	11.8	41.6	31.0	-3.933	<0.001
$KS\ R_x$	0.688	0.274	0.832	0.228	-1.807	0.079
$KS\ R_y$	0.521	0.203	0.705	0.143	-3.324	0.002
$KS\ R_z$	0.652	0.218	0.897	0.136	-4.272	<0.001
$KS\ AU1$	0.413	0.237	0.623	0.165	-3.263	0.003
$KS\ AU2$	0.376	0.224	0.862	0.080	-9.126	<0.001
$KS\ AU4$	0.758	0.241	0.961	0.121	-3.372	0.002
$DTW\ AU1$	115	44.9	128	25.8	-1.105	0.278
$DTW\ AU2$	106	30.3	82.3	26.7	2.604	0.013
$DTW\ AU4$	109	54.8	182	30.8	-5.228	<0.001
$Sync\ AU1$	2.98	3.47	0.00	0.00	3.835	0.001
$Sync\ AU2$	2.33	4.82	0.00	0.00	2.155	0.044
$Sync\ AU4$	15.6	30.4	0.00	0.00	2.293	0.033
$El\ AU1$	173	105	148	88.7	0.817	0.419
$El\ AU2$	222	144	0.00	0.00	6.898	<0.001
$El\ AU4$	365	613	181	170	1.292	0.210

Table 7.5 Mean and standard deviation of agent perception objective measures for the conditions of RA and MM. Significance difference between the condition pair (RA,MM) reported via one-sided Welch's t-test.

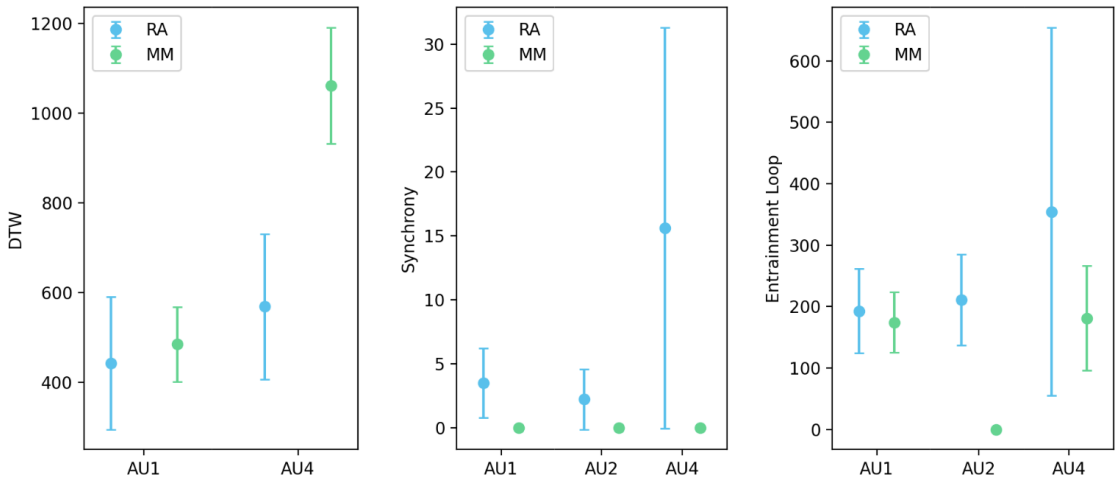


Figure 7.4 DTW resemblance, synchrony, and entrainment loop between adaptive (RA) and non-adaptive (MM) conditions.

For the factors that represent the change of the participant's state before and after (pre & post) the experiment, we calculate their change. For these factors of change (mood, anxiety, and psychological distress), each participant responds to the corresponding questionnaires (mood score, STAI, and K6) twice (before and after the experiment) to measure the change. As the same questionnaires are used twice, we are in the case of having repeated measures. Thus, we can visualize

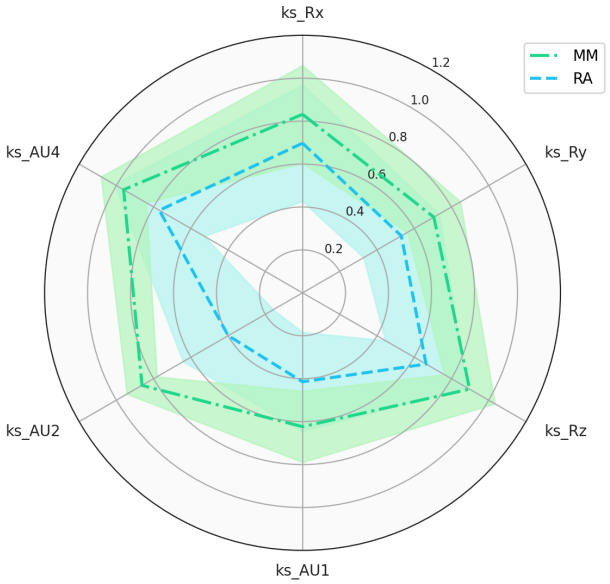


Figure 7.5 KS test between adaptive (RA) and non-adaptive (MM) conditions. The central line in bold represents the mean value and the colored-filled contour represents the standard deviation of each condition.

	RA		MM		SP	
	Mean	SD	Mean	SD	Mean	SD
Mood score						
pre	59.3	24.5	70.0	19.8	51.0	24.1
post	36.0	23.1	35.6	21.9	33.3	22.1
K6						
pre	8.05	3.85	6.50	3.00	7.30	2.75
post	6.10	3.46	5.65	4.72	7.40	3.12
STAI-State						
pre	33.6	9.20	34.5	11.3	35.6	7.98
post	32.3	8.84	33.9	10.5	36.8	9.42
STAI-Trait						
pre	43.5	11.0	42.9	9.86	44.7	7.37

Table 7.6 Mean and standard deviation of user mood and state measures (pre and post) for the conditions of RA, MM, and SP.

the difference to check whether there was a change as shown in Figure 7.6 and Table 7.6.

Mood Change

Looking at the mood scores before (pre) and after (post) the experiment, depicted in Figure 7.6 (left) and Table 7.6, we can remark that the mood score, which

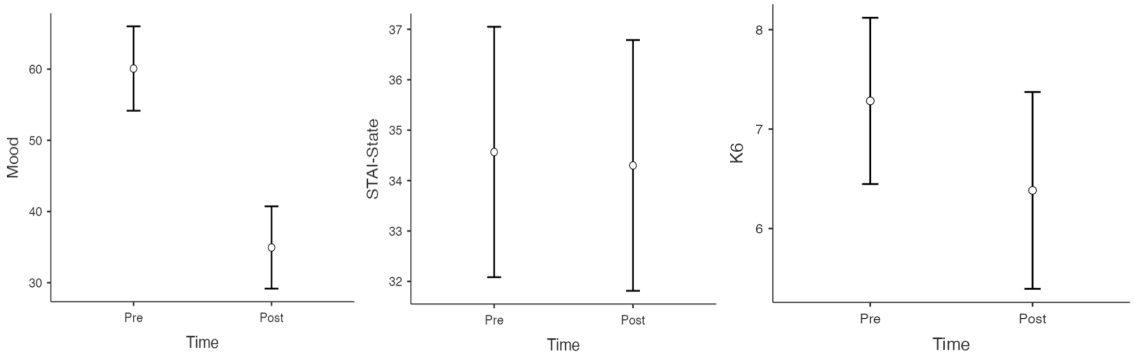


Figure 7.6 Change in user mood (mood scores) and states (anxiety level via STAI-State and psychological distress level via K6) before (pre) and after (post) the experiment.

Measure	RA		MM		SP	
	Mean	SD	Mean	SD	Mean	SD
Mood Change Score	0.376	0.251	0.483	0.313	0.246	0.440
ΔSTAI-State	1.40	7.63	-0.50	4.67	-1.20	6.49
ΔK6	1.95	3.39	0.85	2.74	-0.10	2.43
CC	11.8	7.30	10.2	6.58	11.0	5.20

Measure	Between RA-MM		Between RA-SP	
	Cohen's d	P-value	Cohen's d	P-value
Mood Change Score	-1.192	0.241	1.149	0.260
ΔSTAI-State	0.949	0.350	1.161	0.253
ΔK6	1.128	0.267	2.197	0.035
CC	0.705	0.485	0.374	0.711

Table 7.7 Mean and standard deviation of user mood and state measures for the conditions of RA, MM, and SP. Significance difference between the condition pairs (RA,MM) and (RA,SP) reported via one-sided Welch's t-test.

indicates the participant's emotional state of being depressed, decreases. This indicates that CBT treatment using our agent platform helps participants to change their negative mood to a positive one.

When looking at the mood change scores, shown in Table 7.7, we can remark on the difference in experimental conditions (RA, MM, and SP). Adaptive agent (RA) better ameliorates the participants' mood during the CBT session compared to when the agent retains a still position (SP). However, the non-adaptive condition (MM) presents stronger changes in mood scores. We assume that the presence of the agent's movement (independent of the adaptation) helps in improving the participants' mood as participants may feel at ease by getting the agent's visual feedback. No significance was found between the pairs of (RA,MM) and (RA,SP) for mood change score.

Anxiety State Change

For the anxiety level change, we look into the STAI-State inventory in which participants report their current anxiety level.

The anxiety level obtained by STAI-State, in Figure 7.6 (center) and Table 7.6, shows a slight improvement (decrease in anxiety level) after the experiment (post) compared to that reported before (pre). This may be interpreted as our CBT agent platform also helps the participants to ameliorate their anxiety levels.

The difference in experimental conditions (RA, MM, and SP) shows different effects on the anxiety level change as seen in Table 7.7. Showing reciprocally adaptive agent behavior (RA) seems to help in improving the anxiety state. However, agents with no motion (SP) or showing random ones (MM) may deter the state of the participants by increasing their anxiety level instead of relieving them. No significance was found between the condition pairs for anxiety state change.

Psychological distress level change

By observing the psychological distress level obtained via the K6 scale, shown in Figure 7.6 (right) and Table 7.6, we can see that participants' stress level improves after the experiment (post) compared to that reported before (pre), shown by the decrease in stress levels. This may be interpreted as our CBT agent platform is indeed helpful in ameliorating participants' psychological distress.

When looking at the distress level change, shown in Table 7.7, we can check the difference in experimental conditions (RA, MM, and SP). The psychological distress state seems to improve when the agent is expressive (MM and RA). We can also assume that showing reciprocally adaptive agent behaviors (RA) is more effective in amending the stress level than showing unsynced and random behaviors (MM). When the agent is in a still position (SP), the agent may render the participants in a more stressful state.

Welch's t-test yielded significant variation between the pair of (RA,SP) for distress change while no significance was found for the pair of (RA,MM).

Cognitive change

With the CC scale, results reported in Table 7.7, we can check that the participants all felt the experience of cognitive change and cognitive skill use during the CBT experiment with our agent platform (CC>0). We can also remark that the condition of RA performs better for cognitive change compared to SP and MM conditions. It seems that displaying reciprocally adaptive agent behaviors (RA) helps in experiencing cognitive change and eliciting cognitive skills. We can also check that showing random agent behaviors (MM) rather deters such change. We report no significant statistical differences between the experimental condition pairs.

With this finding, we can validate our hypotheses H1 and H2, showing the relationship of RA>MM and RA>SP respectively, for the CBT effectiveness measures of change in mood, anxiety state, and psychological distress level. Mood change only validates H2 of RA>SP.

7.4.3 Relation between Perception of Agent and User Mood and State Change

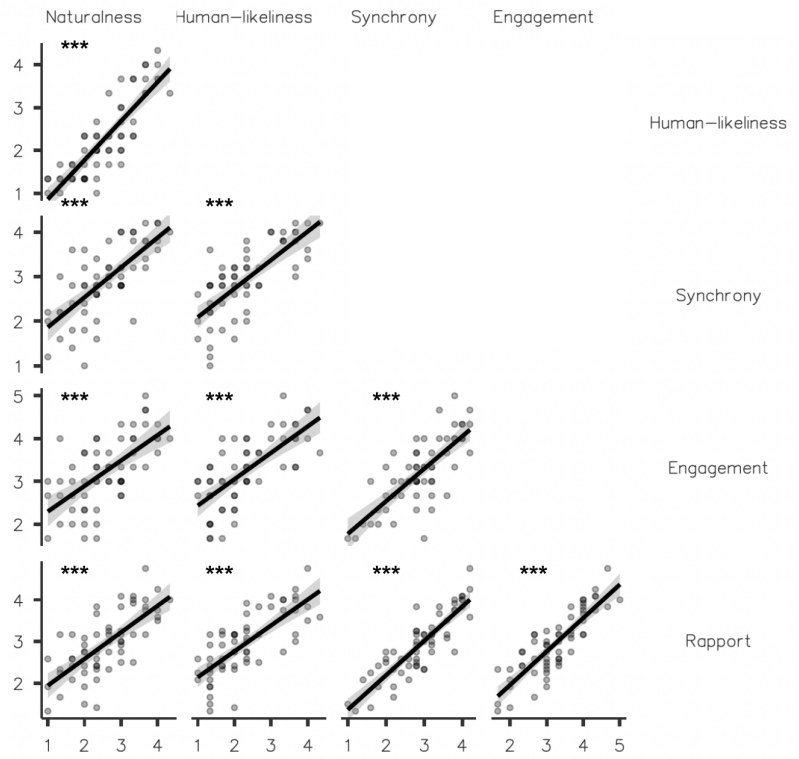


Figure 7.7 Correlation between factors of agent perception with p-values marked as ***:$p < 0.001$ (see Appendix C in Chapter 12 for graph interpretation).

Looking at the scores given by the participants along the factors of the perception of the agent, Figure 7.7, we can note that the 5 factors are heavily correlated (direct relation) to one another with significant differences ($p < 0.001$). We can remark that the factors of agent perception are closely linked with each other. From Figure 7.8, positive correlations between user mood and state change factors are also noticeable, notably between the pairs of (anxiety state change, CC) and (anxiety state change, psychological distress level change). This implies that the mood and state change well complement each other for the participants' CBT performance evaluation.

We are also interested in looking at the relationship between CBT effectiveness (change in mood, anxiety state, distress level, and cognitive state) and the participant's perception of the agent along these 5 factors (naturalness, human-likeliness, synchrony, engagement, rapport). We can remark, via Figure 7.9, that the anxiety level and the 5 factors of the agent's impression are correlated (STAI-State has a direct relation with the 5 factors) with significant differences for naturalness ($p = 0.008$), human-likeliness ($p = 0.049$), synchrony ($p = 0.01$), and rapport ($p = 0.013$). We can observe the same improvement in cognitive change (CC has a direct relation with the 5 factors) differing significantly for all 5 factors at $p < 0.001$. For the other two factors of CBT effectiveness (mood and psychological distress), no significant correlation was found with any of the 5 factors of the

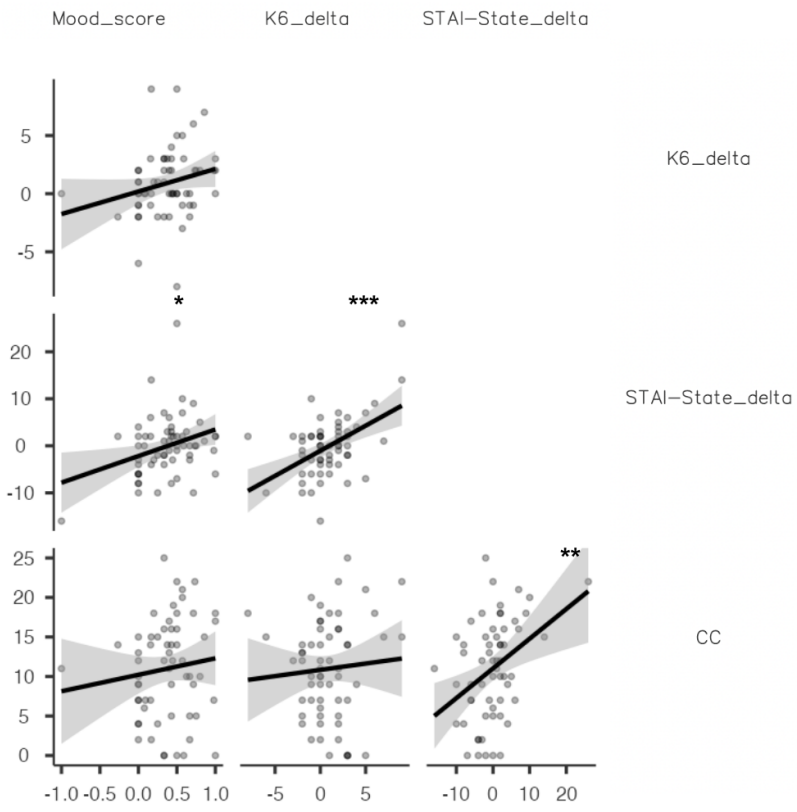


Figure 7.8 Correlation between factors of user mood change and state change with p-values marked as *:$p < 0.05$, **:$p < 0.01$, and ***:$p < 0.001$ (see Appendix C in Chapter 12 for graph interpretation).

perception of the agent. Our results indicate that the impression of virtual agents is positively correlated to user anxiety state and cognitive change.

7.5 Discussion

Through qualitative and quantitative evaluations, we were able to validate that the display of adaptive behavior leads to the *agent's positive perception* of being considered as more *natural*, *human-like*, *in sync*, *engaged*, and having *built a rapport* with the participant. The results are in line with previous studies which demonstrated that agents displaying adaptive behaviors increase the *user's engagement* (Schroder et al. [2011], Ritschel et al. [2017], Weber et al. [2018], Biancardi et al. [2021]), *rapport* (Huang et al. [2010], Raffard et al. [2018]), interaction *synchrony* (Raffard et al. [2018]), and *impression of the agent* (Bailenson and Yee [2005], Huang et al. [2010], Biancardi et al. [2021]). We were able to also remark a non-adaptive agent is perceived less well than a static one (in a still position). This implies that non-adaptive behavior can rather hinder the impression of the agent. Showing just expressive behavior that does not take into account the interlocutor is worse than not showing any behavior.

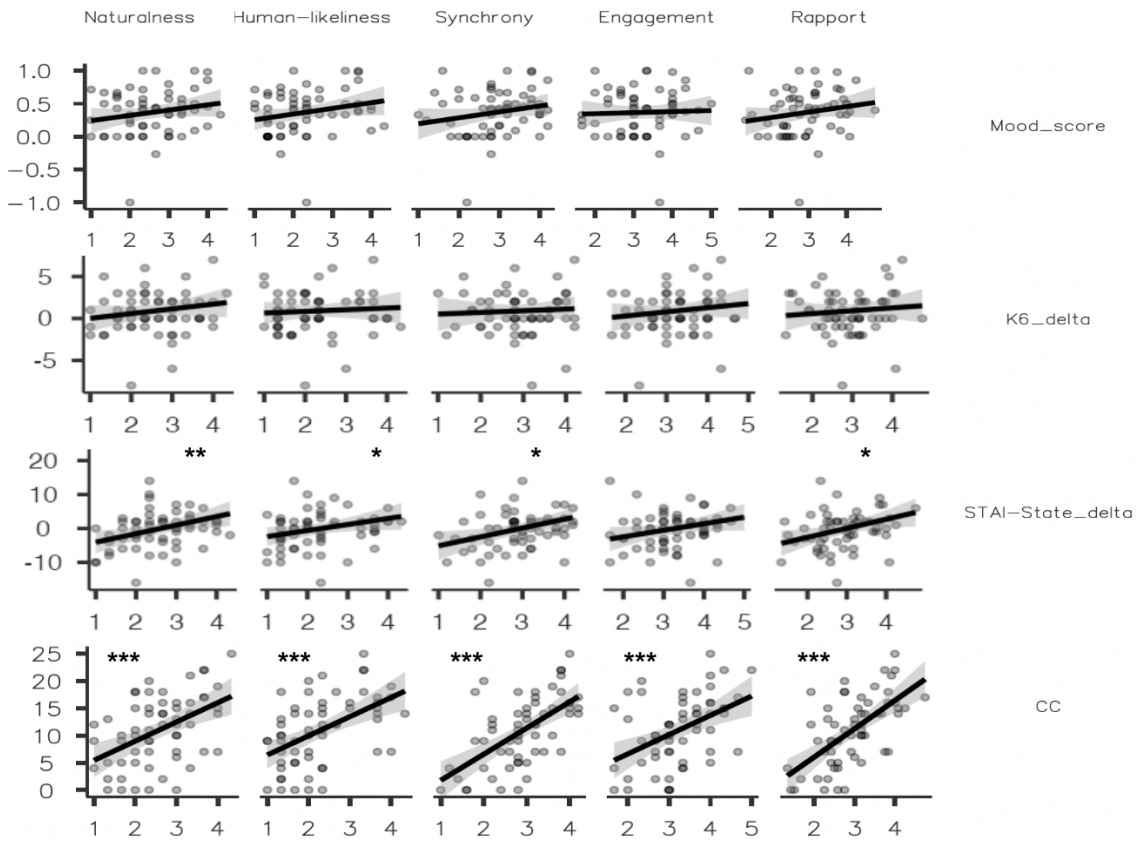


Figure 7.9 Correlation between factors of agent perception, user mood change, and user state change with p-values marked as *:$p < 0.05$, **:$p < 0.01$, and ***:$p < 0.001$ (see Appendix C in Chapter 12 for graph interpretation).

Concerning the *effectiveness* of CBT, we were able to check that our virtual agent system, disregarding the different conditions (RA, MM, and SP), helps deliver CBT in improving users' moods and states of anxiety, psychological distress, and cognitive change. This highlights the advantage of deploying virtual agents in general notably in healthcare. The adaptive condition showed its usefulness in having higher effects of CBT. We noticed that agents showing random expressive behavior can still improve the impact of the application. These results are consistent with findings in the medical and psychological fields that rapport (VandenBos [2007]) affects the effectiveness of CBT (Asay and Lambert [1999], Ardito and Rabellino [2011], Norcross and Lambert [2018]), and in the engineering field that nonverbal behavior affects rapport (DeVault et al. [2014], Huang et al. [2010], Raffard et al. [2018]) as defined in the context of human-agent interaction (Gratch and Lucas [2021]). This may be related to the presence of the agent that the participants feel during the CBT session. Expressive agents may heighten the feeling of the agent being physically present with the participants thus leading to a psychological ease to interact and benefit the effect of the interaction and thus the CBT treatment.

We also found correlations between the factors of perception of the agent's impression, between the user's mood and states, and between the perception of the

agent's impression and the user's anxiety state and cognitive change. The positive correlation observed between the factors of perception of the agent's impression, revealing a halo effect, shows that the factors are indeed tightly connected which can be grouped as a global impression of the agent. The correlations found between user mood and state change factors, especially the positive relation between anxiety and distress state changes, denote that these factors complement one another in showing the effect of CBT. The interdependence seen between agent perception and user states (anxiety and cognitive) may be interpreted as the user states can be ameliorated by improving the agent's impression which can be done by endowing the agent with the skill to adapt.

7.6 Social Skills Training System Application

Our *real-time* adaptive virtual agent system proved its usefulness for the proof-of-concept of CBT. To manifest that our system is applicable to various domains, its use needs to be confirmed with other use cases. To address this, we applied our system to Social Skills Training (SST; Bellack et al. [2013]), a behavioral therapy for improving social skills in people, to demonstrate our system's applicability and effectiveness to other usages.

Our system applied for SST (Saga et al. [2023b]) replaces the CBT-related module (automatic thought classifier) with the SST-related module. The SST-related module is based on the system of Saga et al. [2023a], comprised of:

- SST evaluation module: estimates eye contact, facial expression, and vocal variation scores ranging from 1 to 5. The scores are predicted via random forest models based on multimodal features (average voice intensity, $F0$, smile, head poses, nodding, facial AUs, and gestures.;
- SST feedback module: selects a set of pre-defined SST performance feedback sentences to reflect users' nonverbal behaviors during the interaction.

We are interested in whether participants interacting with our system perceive the agent in a similar way for different applications, in our case the SST and CBT. For the SST experiment, we followed the same experimental protocol and questionnaires of human-agent interaction evaluation as in the CBT experiment. Furthermore, the SST experiment is conducted under the reciprocal adaptation (RA) condition. 15 French-speaking participants with similar demographics, as in the CBT experiment, were recruited. We compare the agent perception results of the two along the 4 factors of naturalness, human-likeliness, synchrony, and engagement.

We check, in Figure 7.10 and Table 7.8, that for both applications of SST and CBT, displaying reciprocally adaptive behavior, the agent is perceived in the same way in terms of *naturalness*, *human-likeliness*, *synchrony*, and *engagement*. No significance was found for all aspects between CBT and SST applications.

We have also asked participants to respond to additional questionnaires concerning the SST effectiveness evaluation. The analysis of the effect of SST is ongoing. Like CBT, we expect to further improve the SST performance with our adaptive agent.

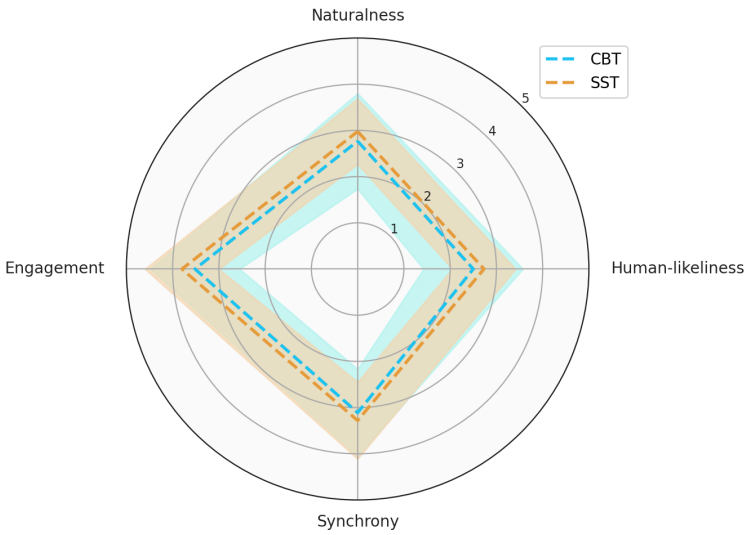


Figure 7.10 Perception of the agent's reciprocally adaptive behavior of SST and CBT applications (RA condition for both) along naturalness, human-likeliness, synchrony, and engagement. The central line in bold represents the mean value of each application (SST and CBT) and the colored-filled contour represents the standard deviation of each condition.

Factor	CBT		SST	
	Mean	SD	Mean	SD
Naturalness	2.77	1.030	2.98	0.707
Human-likeliness	2.50	1.070	2.73	0.681
Synchrony	3.11	0.941	3.28	0.831
Engagement	3.52	0.964	3.80	0.795

Factor	Between CBT-SST	
	Cohen's d	P-value
Naturalness	-0.68	0.501
Human-likeliness	-0.74	0.467
Synchrony	-0.56	0.582
Engagement	-0.93	0.362

Table 7.8 Mean and standard deviation of agent behavior perception factors for the applications of SST and CBT. Significance difference between the application pairs (SST,CBT) reported via one-sided Welch's t-test.

7.7 Contributions and Conclusion

7.7.1 Contributions

Our work makes the following contributions:

- We propose an adaptive SIA system that provides *real-time* human-agent interaction;

- The display of adaptive SIA behavior can further improve the user experience in terms of the user perception of the *agent's impression* (naturalness, human-likeliness, synchrony, engagement, and rapport);
- Adaptive agent is *effective* in delivering CBT as it can enhance users' moods and states (anxiety, psychological distress, and cognitive).
- Adaptive agent gives a similar *impression* (naturalness, human-likeliness, synchrony, and engagement) for different applications of CBT and SST.
- A human-agent interaction database (*CBT-HAI DB*) has been collected by recording the CBT interactions and made available to the research community (available after signing the EULA form).

7.7.2 Conclusion

Adapting to other people is an essential communication skill. Virtual agents interacting with their users need to know how to adapt to their interlocutors to provide a lively and interesting interaction. In this work, we developed a virtual agent endowed with adaptation capacity that is capable of functioning in *real-time*. The healthcare application of CBT was selected as a proof-of-concept to demonstrate the use of our adaptive agent. The agent showed its utility in delivering CBT by enhancing users' moods and states (anxiety, psychological distress, and cognitive). Moreover, showing adaptive behavior can further improve the *user experience* in terms of the user perception of the a*gent's impression* along the factors of *naturalness*, *human-likeliness*, *synchrony*, *engagement*, and *rapport*. However, expressive agents are not always positively perceived since non-adaptive ones rather deter the user's impression. Through our study, expressive agents (adaptive and non-adaptive) have shown their *effectiveness* in improving users' negative moods and states, compared to static ones (in a still position). Adaptive agents are full of promise as they can ameliorate *user impressions* and be employed for various applications.

The key points of this Chapter:

Addressing Hypotheses

- Adaptive and expressive agents can further enhance the *user experience* (*agent impression*) and the *effectiveness* of CBT (user mood and state change).
- Agents enabled with reciprocal adaptation (RA) are expected to outperform agents showing random behaviors (or mismatched behavior; MM) and still agents (or in a still position; SP).

Real-time Adaptive SIA System

- An adaptive SIA system functioning in *real-time* and delivering CBT treatment.
- The display of adaptive SIA behavior enhances the *user perception* of the *agent's impression* (*user experience*) in terms of *naturalness*, *human-likeliness*, *synchrony*, *engagement*, and *rapport*;
- CBT becomes *effective* with an adaptive agent compared to a non-expressive and non-adaptive agent. It improves users' moods and states (anxiety, psychological distress, and cognitive).
- Application of our adaptive SIA system to SST. A similar agent impression is given to the users for different applications (SST and CBT).
- Novel human-agent interaction database (*CBT-HAI DB*).

Publications

- *(Under revision, Submitted to IJHCS)* - Jieyeon Woo, Kazuhiro Shidara, Catherine Achard, Hiroki Tanaka, Satoshi Nakamura, and Catherine Pelachaud. Adaptive virtual agent: Design and evaluation for real-time human-agent interaction. *International Journal of Human-Computer Studies*, 2023f
- Jieyeon Woo, Michele Grimaldi, Catherine Pelachaud, and Catherine Achard. Iava: Interactive and adaptive virtual agent. In *ACM International Conference on Intelligent Virtual Agents (IVA '23)*, 2023c
- Jieyeon Woo, Michele Grimaldi, Catherine Pelachaud, and Catherine Achard. Conducting cognitive behavioral therapy with an adaptive virtual agent. In *ACM International Conference on Intelligent Virtual Agents (IVA '23)*, 2023b
- Takeshi Saga, Jieyeon Woo, Alexis Gerard, Hiroki Tanaka, Catherine Achard, Satoshi Nakamura, and Catherine Pelachaud. An adaptive virtual agent platform for automated social skills training. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2023b

Part V

Modeling Reciprocal Adaptation with Intrapersonal Memory

Chapter 8

Modeling Reciprocal Adaptation with Intrapersonal Memory

Contents

This Chapter presents the *HI²-ADAM* model that generates *adaptive SIA* behavior as both *speaker* and *listener* by encoding the *multimodality* and *intrapersonal* (explicit modeling of *modality histories*) and *interpersonal relationships*. The deeper encoding of *multimodality* and explicit modeling of the *intrapersonal* and *interpersonal temporalities* show promising results in rendering *SIA* behavior that performs well in terms of *reciprocal adaptation resemblance* (and interlocutor *synchrony*) and behavior *human-likeliness*.

8.1 Introduction

In Chapter 6, we were able to generate reciprocally adaptive *SIA* behavior via the *ASAP* model which modelizes the *interpersonal relationship* (or *reciprocal adaptation*). The *ASAP* model notably improved the agent's *interpersonal dynamics* of synchrony and engagement.

The behavior adaptation shown during an interaction is *interpersonal*, interlocutors adapt to each other, but the behaviors are also coordinated *intrapersonally*. To better generate adaptive behavior, unlike the *ASAP* model which only models the *interpersonal relationship*, in this chapter we also modelize the *intrapersonal relationship*. We propose a new model, *HI²-ADAM*, which explicitly captures *intrapersonal relationship* by modeling each *modality history* (or *memory*) of each interlocutor and learn from the relation between these different *histories*. It also has a deeper encoding of *interpersonal relationship* present between the interlocutors.

We expect that the better management of *multimodality* of *HI²-ADAM* can enhance the *behavioral aspects* of naturalness and human-likeliness. This management could be done by providing explicit modeling of the *multimodality* within the *intrapersonal* relationship notably by modeling each *modality history* (or *memory*) and the relation between the different histories.

Our overall aim is to create a *social* and *engaging SIA* by modeling its *behavior adaptation* (explicitly capturing *intrapersonal* and *interpersonal relationships* and *multimodality*) while ensuring *behavior continuity*.

We pose the following research questions (RQs):

- RQ 1: *interpersonal dynamics* of synchrony and engagement can be augmented by modeling the *interpersonal relationship* (or *reciprocal adaptation*);
- RQ 2: *behavioral aspects* of naturalness and human-likeliness can be improved by capturing the *intrapersonal relationship* with the modelization of the relation between *modality histories*.

We propose ***Historical intrapersonal interpersonal ADAptive Multimodal (HI²-ADAM)*** model, a novel method to synthesize *adaptive facial* gesturing for *SIAs*. We explicitly model the *intrapersonal relationship* by encoding the prior emitted *multimodal* signals and their *histories* (*modality memory*) while ensuring *motion continuity*. We model the *interpersonal relationship* (or *reciprocal adaptation*) from the learned *intrapersonal* representation encodings to generate *SIA* behavior for both roles of *speaker* and *listener*. We explore the best way to capture the *reciprocal adaptation* to generate *adaptive* nonverbal *SIA* behavior within a dyadic setting. *intrapersonal* and *interpersonal relations* are learned through *attention mechanisms*.

The chapter is organized as follows. Section 8.2 describes the proposed *HI²-ADAM* model architecture and reports our experiments (objective and subjective evaluation results and discussions). We finally report our contributions and conclude in Section 8.3.

8.2 Historical intrapersonal interpersonal ADAptive Multimodal (*HI²-ADAM*) Model

8.2.1 Model Architecture

Like the *ASAP* model (ref. Chapter 6), we focus on generating *adaptive* nonverbal *SIA* behavior as both a *speaker* and *listener*. We train on real human-human interactions, to learn human-human *interpersonal* and *intrapersonal relationships* for *SIA* and simulate our predictions on a *SIA*.

We propose a new model architecture, the Historical intrapersonal interpersonal ADAptive Multimodal model (*HI²-ADAM*), to synthesize *adaptive facial gestures* for *SIA*s. It takes as input *speech* and *facial gestures* of both *SIA* (A) and *User* (U), corresponding to their past behaviors (behaviors they displayed so far), and predicts *SIA*'s and *User*'s *facial gestures* at the next time step. We choose to synthesize the gestures of both *SIA* and *User*, during the training phase, to better learn *interpersonal* and *intrapersonal relations*.

We employ similar features as for the *ASAP* model (ref. Chapter 6) which are composed of facial gestures ($G_{x,y}$, $R_{x,y,z}$, and *AUs* (1,2,4,6, and 12)) and speech features ($F0$, loudness, voicing probability, *MFCCs* (0-12)).

HI²-ADAM model operates as follows. It takes as input the 100 past frames ($t-99:t$; found through empirical tuning and also used in Woo et al. [2021] and in *ASAP* (ref. Chapter 6)), where t is the current frame, of the:

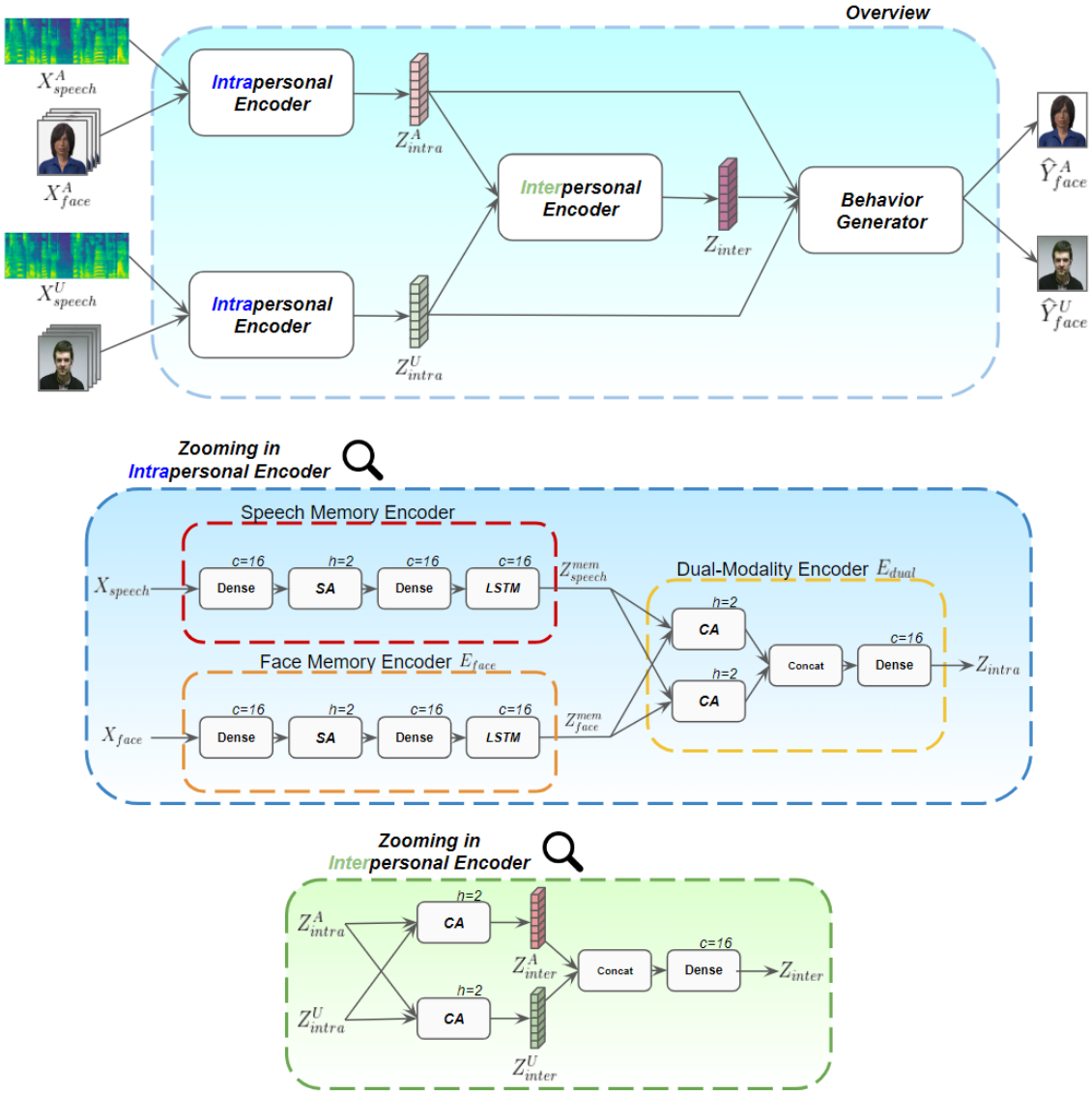
1. *Speech features* of A (X^A_{speech}) and those of U (X^U_{speech}),
2. *Facial features* of A (X^A_{face}) and those of U (X^U_{face}).

For each prediction of the next frame ($t+1$), the model predicts:

1. A's *facial gestures* ($\widehat{Y}^A_{face}$),
2. U's *facial gestures* ($\widehat{Y}^U_{face}$).

HI²-ADAM consists of three main components, as illustrated in Figure 8.1. The first component is the intrapersonal encoder E_{intra}, which explicitly encodes the *intrapersonal relation* via the management of multimodal signals. It manages the multimodality and learns the *intrapersonal relation* from each *modality history* via a *modality memory schema*. This schema consists of encoding each *modality* - *speech features* and *facial features* - corresponding to the past 100 frames. The second component is the interpersonal encoder E_{inter}, which encodes the *interpersonal relations* by applying cross-attentions between *A*'s and *U*'s features' embeddings. The last component is the *behavior generator* which generates *A*'s and *U*'s *facial gestures* of the next frame. These components are detailed in the following.

Intrapersonal Encoder (E_{intra}) As shown in Figure 8.1, E_{intra} takes as input X_{speech} and X_{face} of either *A* or *U* and generates the corresponding intrapersonal embedding Z_{intra}. It consists of two sub-encoders. The first is the modality memory encoder (E_{speech} or E_{face}). The second is the dual-modality encoder E_{dual}.



HI²-ADAM model architecture

Figure 8.1 ***HI²-ADAM*** (**H**istorical **I**ntra-personal **I**nter-personal **ADA**ptive **M**ultimodal) model architecture. The intrapersonal encoder (E_{intra}) takes the *speech* X_{speech} and the *facial gestures* X_{face} of the previous 100 frames of either the *SIA* (A) or the *User* (U) to encode the corresponding *intrapersonal relationship* Z_{intra}. The interpersonal encoder (E_{inter}) learns from *intrapersonal relationships* Z^A_{intra} and Z^U_{intra} to encode the *interpersonal relationship* between them Z_{inter}. The behavior generator (G_{face}) takes Z^A_{intra}, Z^U_{intra}, and Z_{inter} to generate the sequence of *facial gesture* for the next frame at $t+1$ $\widehat{Y}^A_{face}$ and $\widehat{Y}^U_{face}$. At *training time*, *HI²-ADAM* is trained with human-human (U_1-U_2) interactions (U_1 for A and U_2 for U) and predicts both of humans' facial gestures ($\widehat{Y}^{U_1}_{face}$ and $\widehat{Y}^{U_2}_{face}$). At inference time, *HI²-ADAM* renders the facial gestures of A and U. To infer the next A's behavior, we feed back the predicted A's behavior and the ground truth of U.

Modality Memory Encoder (E_{speech} or E_{face}) Both E_{speech} and E_{face} takes its corresponding *modality* - X_{speech} or X_{face} respectively - as input and renders the

modality memory embedding Z^{mem}_{speech} and Z^{mem}_{face} representing the past 100 frames. Each corresponding modality memory encoder firstly learns the *modality specific* information by applying *self-attention*, with a $h = 2$ (where h is the head size), preceded and followed by dense layers (E_d) with $c = 16$ (where c is the cell size). Then, it embeds the memory sequence of the chosen *modality* via a LSTM layer (E_m) with $c = 16$, as depicted in Figure 8.1. It takes X_{mod} - where mod represents either *speech* or *facial gestures* - and outputs Z^{mem}_{mod}, and can be expressed as:

$$Z^{mem}_{mod} = E_m\left(E_d\left(SA\left(E_d\left(X_{mod}\right)\right)\right)\right) \tag{8.1}$$

where $SA(\cdot)$ denotes self-attention layer.

Dual-modality Encoder (E_{dual}) E_{dual} captures the relationship between the multimodal signals by applying *cross-attention mechanisms* on the corresponding modalities: CA^{speech} and CA^{face} with $h = 2$ followed by E_d with $c = 16$, as shown in Figure 8.1. CA^{speech} has a query Q equals to Z^{mem}_{speech} with key K and value V equal to Z^{mem}_{face}. CA^{face} has a query Q equals to Z^{mem}_{face} with key K and value V equal to Z^{mem}_{speech}. E_{dual} takes Z^{mem}_{speech} and Z^{mem}_{face} as inputs and generates Z_{intra}. It can be written as:

$$\begin{aligned} Z_{intra} = E_d\left(\left[\, CA_{speech}\left(Q_{speech}, K_{face}, V_{face}\right),\right.\right. \\ \left.\left. CA_{face}\left(Q_{face}, K_{speech}, V_{speech}\right)\,\right]\right) \end{aligned} \tag{8.2}$$

where $CA(Q, K, V)$ denotes cross-attention layer and $[\cdot]$ denotes concatenation layer.

Interpersonal Encoder (E_{inter})

As illustrated in Figure 8.1, E_{inter} takes as input Z^{A}_{intra} and Z^{U}_{intra}, which are the *intrapersonal* representations of A and U respectively. It renders Z_{inter}, a representation of *interpersonal relation* between A and U. E_{inter} applies *cross-attention mechanisms* on the both *intrapersonal* representations: CA^{A} and CA^{U} with $h = 2$ followed by E_d with $c = 16$. CA^{A} has a query Q equals to Z^{A}_{intra} with key K and value V equal to Z^{U}_{intra}. CA^{U} has a query Q equals to Z^{U}_{intra} with key K and value V equal to Z^{A}_{intra}. It can be written as:

$$Z_{inter} = E_d\left(\left[CA^{A}\left(Q^{A}, K^{U}, V^{U}\right), CA^{U}\left(Q^{U}, K^{A}, V^{A}\right)\right]\right) \tag{8.3}$$

where $CA(Q, K, V)$ denotes cross-attention layer and $[\cdot]$ denotes concatenation layer.

Behavior Generator (G_{face})

G_{face} takes as input the:

1. A's *intrapersonal* representation (Z^{A}_{intra}),
2. U's *intrapersonal* representation (Z^{U}_{intra}),
3. *interpersonal* representation of A and U (Z_{inter}).

It generates the corresponding *facial gestures* $\widehat{Y}^{A}_{face}$ and $\widehat{Y}^{U}_{face}$ by decoding with a dense layer (D_d) with $c = 20$, as depicted in Figure 8.1. The final outputs $\widehat{Y}^{P}_{face}$ can be written as:

$$\widehat{Y}^{P}_{face} = D_d\left(Z^{P}_{intra}, Z_{inter}\right) \tag{8.4}$$

where P represents either *A* or *U*. At training time, *HI²-ADAM* synthesizes $\widehat{Y}^{U}_{face}$ to better learn *interpersonal* and *intrapersonal relations*. $\widehat{Y}^{U}_{face}$ is disregarded at inference time since the aim is to predict only A.

8.2.2 Implementation Details Training Regime

We train our model on real human-human (U_1-U_2) interactions using the French *NoXi* dataset (Cafaro et al. [2017]) as explained in Chapter 4. The *HI²-ADAM* model uses the same features as the *ASAP* model presented in Chapter 6. The employed features consist of visual (G_x, G_y, R_x, R_y, R_z, $AU1$, $AU2$, $AU4$, $AU6$, and $AU12$) and acoustic features (fundamental frequency, loudness, voicing probability, and 13 MFCCs). Our model learns to synthesize adapted gestures of U_1 and U_2. During inference, *HI²-ADAM* synthesizes the behavior of A and U. $\widehat{Y}^{A}_{face}$ is inferred using the previous prediction of A and the ground truth of U. We apply *adaptive online prediction* to generate *continuous A*'s behavior in an *autoregressive* fashion.

We split our dataset into 3 sets: training (70%), validation (10%), and test (20%). The test set does not include data of *speakers* and *listeners* that are seen during training. The aim is to test *HI²-ADAM*'s capacity to extrapolate on new unseen *speakers* and *listeners* and therefore its capability to generalize.

To train our model, we use the Mean Squared Error (MSE) as our loss function and the Adam optimizer (Kingma and Ba [2014]) with Cyclical Learning Rate (CLR) (Smith [2017]) (*triangular* learning rate policy, *base_lr* of $1e-7$, *max_lr* of $1e-3$, and step size factor of 10). The training was done for 300 epochs (with an average runtime of $125h$) on a $2.2GHz$ Intel Xeon Linux server with NVIDIA GeForce GTX TITAN X and $64GB$ RAM with a batch size of 32. The best set of hyperparameters is chosen after manual optimization, via manual grid search, based on the validation set.

8.2.3 Baselines

For the evaluation, we compare *HI²-ADAM* against the *baseline models (base)* which learns from the *interpersonal relationship*. The baselines are as follows.

- **IL-LSTM (Dermouche and Pelachaud [2019b])**: generates *SIA*'s facial gestures based on unimodal input features (facial gestures) of its own (*A*) and those of the human user (*U*). A LSTM model with a sliding window prediction is used making it prone to jerky movements.
- **Symmetrized IL-LSTM with online LSTM (sym-IL-LSTM**; Woo et al. [2021]**)**: generates *SIA*'s facial gestures by modeling the *multimodality* of *speech* and *facial gestures* of both itself (*A*) and the human user (*U*), and assures *motion continuity* by employing a LSTM with adaptive online prediction and autoregression.

- ***ASAP* (ref. Chapter 6)**: generates reciprocally adaptive and continuous *SIA*'s facial gestures learned from *speech* and *facial gestures* of both itself (*A*) and the human user (*U*). The *interpersonal relationship* and *multimodality* are learned via a LSTM with attention mechanism of transformers and pruning technique. Furthermore, the *motion continuity* is assured via the autoregressive adaptive online prediction. The *intrapersonal relationship* and its history is not modelized.

8.2.4 Objective Evaluation

To assess our model, we conduct an objective evaluation to check its performance against the state-of-the-art approaches, which we select as our baselines, and to verify the effectiveness of each *HI²-ADAM*'s key components through ablation studies.

Objective Evaluation Measures

We want to assess whether the generated behavior is *appropriate* and *reciprocally adaptive*. To do so, we employ the metrics used for *ASAP* (ref. Chapter 6) of RMSE and Kolmogorov-Smirnov two-sample test (KS test) (Massey Jr [1951]) to measure the *behavior appropriateness* of A's predictions ($\widehat{A}$) against its ground truth (GT) behavior (A). We also employ the DTW (Müller [2007]) resemblance for *reciprocal adaptation resemblance* assessment.

In addition to these metrics, **MAE** is used to measure the distance between the predictions and GT to measure the generated error along with RMSE. To thoroughly evaluate the *reciprocal adaptation resemblance*, of which we measure the resemblance between $\widehat{A}$ and U's GT data (U), new metrics are performed, in addition to DTW resemblance. The additional *reciprocal adaptation resemblance* metrics are as follows:

- **Time lagged cross-correlation coefficient (TLCC)** (Boker et al. [2002]): linear relationship invariant to speed, which is used to quantify *global synchrony*. TLCC is computed in chunks of $8s$ with a time lag of $2s$ as in Ng et al. [2022].
- **Synchrony (Sync) and Entrainment Loop (EL)** (ref. Chapter 5): synchrony and entrainment loop measures to evaluate the *reciprocal adaptation* between $\widehat{A}$ and *U*.

Lower values denote better performance for MAE, RMSE, and KS test. For the *resemblance metrics* (TLCC, DTW, Sync, and EL), the closer the value of the metric is to the GT, the better the model performs in generating *adaptive A*'s behaviors.

Objective Evaluation Results and Discussion

The evaluation results are listed in Table 8.1. Δ_{base} represents the change in performance over the best-performing baseline approach for each metric.

	MAE	RMSE	KS test	TLCC	DTW	Sync	EL
GT				0.334	1317.5	132.4	1172.3
IL-LSTM	0.304	0.415	0.329	0.343	1216.2	45.3	323.3
sym-IL-LSTM	0.180	0.227	0.284	**0.335**	1281.3	33.3	232.3
ASAP	0.185	0.254	**0.282**	0.317	1399.3	142.0	1890.5
HI²-ADAM-noE$_m$	0.099	0.132	0.515	0.271	1228.4	79.1	603.8
HI²-ADAM-noE$_{dual}$	0.143	0.186	0.396	0.300	1352.6	262.3	2255.8
HI²-ADAM-noE$_{inter}$	0.136	0.178	0.406	0.261	1127.8	82.3	586.3
HI²-ADAM	**0.156**	**0.197**	0.437	0.291	**1319.6**	**137.4**	**989.0**
Δ_{base}	↓ **0.024**	↓ **0.030**	↑ 0.155	↑ 0.042	↓ **79.7**	↓ **4.6**	↓ **534.9**

Table 8.1 Objective evaluation of HI²-ADAM against the baselines along with ablations using the selected metrics. GT denotes ground truth interaction. The best results are highlighted in **bold**. Δ_{base} represents the change in performance over the best-performing baseline approach of each metric. Δ_{base} entries in green when HI²-ADAM outperforms best baseline, in red when it is not the case.

Comparing with Baselines We remark that *HI²-ADAM* outperforms the *baselines* in terms of *behavior appropriateness*. This is reflected through low errors of MAE and RMSE represented by Δ_{base} ($\downarrow 0.024$ and $\downarrow 0.020$ respectively). For the density distribution, via the KS test, we observe that *HI²-ADAM* performs comparatively less than the *baselines* indicating that *HI²-ADAM* possesses the least similar density distribution compared to that of the GT. In detail, *ASAP* performs the best in terms of having the most similar density w.r.t. GT (0.282) and *HI²-ADAM* the worst (0.437) with Δ_{base} of $\uparrow 0.155$. This low performance of *HI²-ADAM* does not imply that it generates wrong *SIA* behavior but that it has either a smaller or a wider range of behavior variety than that of the GT. The focus of this study is not to produce a variety of behaviors but to generate *SIA* behaviors that are *adaptive* to its interlocutor. Thus, this weak performance of the KS test metric is not critical for our aim. Moreover, *HI²-ADAM* performs the best in terms of the *reciprocal adaptation resemblance* metrics as seen in the Table 8.1. DTW, synchrony, and entrainment loop of $\widehat{A}\&U$ show that *HI²-ADAM* resembles the GT the most with Δ_{base} of $\downarrow 79.7$, $\downarrow 4.6$, and $\downarrow 534.9$ respectively. For the three measures of DTW, synchrony, and entrainment loop, we report the trend for smile (i.e. AU12), as it is a very important social signal, and as the synchrony measures (Sync and EL) were proposed for AU12. For the other AUs, we observed that the proposed measure significantly outperformed the state-of-the-art methods for eyebrow movements (AU1 and AU2). With TLCC, we remark that the sym-IL-LSTM is the closest to the GT while *HI²-ADAM* is the farthest one with Δ_{base} of $\uparrow 0.042$. As DTW considers the variation of sequence length while being invariant to speed unlike TLCC, it represents better the *global correlation*. Thus, for the interpretation, we can put more emphasis on the DTW results compared to that of TLCC. This comparative study shows that the inclusion of explicit modeling of *intrapersonal relation* leverages the quality of produced gestures in terms of both *behavior appropriateness* and *reciprocal adaptation resemblance*.

Ablation studies To check for the effectiveness and influence of each of *HI²-ADAM*'s key encoders, we conduct additional ablation studies. We perform the ablations of:

- Modality Memory Encoder E_m (noE_m),
- Dual-modality Encoder E_{dual} (noE_{dual}),
- Interpersonal Encoder E_{inter} (noE_{inter}).

The ablation of each of *HI²-ADAM* key encoder - E_m, E_{dual}, and E_{inter} - results in the improvement of the *reciprocal adaptation resemblance*. This is seen by an increase in DTW (87.0, 33.0, and 187.6 respectively), synchrony (48.3, 124.9, and 45.1 respectively), and entrainment loop resemblance (385.2, 900.2, and 402.7 respectively). TLCC shows that the insertion of E_m improves the *HI²-ADAM* by 0.020 along with E_{inter} by 0.030. However, E_{dual} slightly deteriorates the performance by 0.009. As in the *baseline comparison study*, it is better to concentrate on the other *reciprocal adaptation resemblance* metrics as DTW is a more dynamic measure of synchrony than TLCC. However, this enhancement of *reciprocal adaptation resemblance* is at the expense of lowering its *behavior appropriateness* performance. This is observed via MAE (0.057, 0.013, 0.020 respectively) and RMSE (0.065, 0.011, 0.019 respectively). The same conclusion can be drawn by looking at KS test result. The fall of performance is seen for the additions of E_{dual} (0.041) and E_{inter} (0.031) while E_m improves (0.078). This compromise of losing *behavior appropriateness* to gain an *adaptive* one may be a good exchange. It is more valuable to generate *SIA* behaviors with *adaptation capacity* than to reproduce the same GT behavior. In fact, in a human-human interaction, there could be multiple possible behaviors and generation timings facing the same interacting partner's behavior. This might vary depending on the various factors such as the context, situation, and interlocutor's personality and mood.

We can conclude that it is important to model the *intrapersonal relation* with the encodings of E_m and E_{dual}, and the *interpersonal relation* with E_{inter} to synthesize *adaptive* nonverbal *facial gestures* for both roles as *speaker* and *listener*.

8.2.5 Subjective Evaluation

To fully evaluate the perceived quality of the generated agent's behavior, we conduct a subjective evaluation via user perceptive study. The user study complements the objective evaluation by looking into the influence of our *HI²-ADAM* model on the perception of the aspects of: 1) the generated agent behaviors' naturalness and human-likeliness; 2) the *interpersonal dynamics* such as the synchrony between the interlocutors and their engagement.

The perception of human-agent interaction is evaluated along the constructs of behavior naturalness, behavior human-likeliness, interaction synchrony, and engagement using the same questionnaires as in Chapter 6.

Subjective Evaluation Method

The perceptive study is conducted on Prolific, an online crowd-sourcing platform. 8 video clips, each with a duration of $10s$, extracted from the NoXi database are used. The video clips consist of a *SIA* and a human participant conversing while taking speaking turns. They have both speaking and listening turns. A participant that has the speaking turn talks about a common topic and the other that has the listening turn reacts via nonverbal behaviors (visual and acoustic) including backchannels (head nod and/or verbal reply of "yes").

For the subjective evaluation, we consider 4 conditions: IL-LSTM, *ASAP*, *HI²-ADAM*, and GT (ground truth human-human interaction from NoXi). We replace one of the human participants with a *SIA*. The *SIA* keeps the speech of the human participant that it is replacing, but its facial expressions and its head movements are driven by one of the computational models (IL-LSTM, *ASAP*, and *HI²-ADAM*). For the ground truth condition, we use the same evaluation setting of visualizing it on the *SIA* for the fairness of quality visualization. This choice of keeping the same setting is to eliminate any impact that may be caused by the participants' impression of the virtual character (Shiban et al. [2015]).

For the *SIA* animation generation, the Greta platform (Niewiadomski et al. [2009]), an open source *SIA* platform, is used. The *SIA*'s visual animation is merged with the audio of the ground truth. For the perception study, we display a *SIA* and a human participant side-by-side, as seen in Figure 8.2. As we do not render the mouth movements, we blur them so that they won't hinder the evaluators' perception of the *SIA* during the assessment as for *ASAP*.

For each of the 8 video clips extracted from NoXi, 4 videos of the *SIA* are created. Each of the 4 videos displays the *SIA* behavior of either one of the computational models (IL-LSTM, *ASAP* or *HI²-ADAM*) or the ground truth. A total of 32 ($8 * 4$) human-agent videos, as shown in Figure 8.2, are used as stimuli in our study.

Figure 8.2 Example of a video clip used for the user perception study. It depicts an interaction between a *SIA* (left) and a human participant (right).

To lighten the workload of our crowd-sourced users, as a long evaluation may lower the users' concentration and hinder the perception study, we split the study into two groups. Each group evaluates 16 videos of 10 seconds (4 conditions for each of the 4 video clips) and the videos are shuffled randomly to circumvent the

impact that the ordering might have on the study. 60 participants are recruited for the group study and evaluate each video (16 videos per group) with the aforesaid set of 14 questions to evaluate the 4 constructs. We filter out untrustworthy crowd-sourced works via attention check questions (e.g. "Is the virtual character swimming?").

Subjective Evaluation Results and Discussion

For the analysis, we group the participants' responses of the aforementioned 14 questions into the 4 constructs of behavior: naturalness, behavior human-likeliness, synchrony, and engagement. The distribution of each condition, which are GT, IL-LSTM (Dermouche and Pelachaud [2019b]), *ASAP* (ref. Chapter 6), and *HI²-ADAM*, is represented for each of the 4 constructs, Figure 8.3 and 8.4. We also report the median and mean values in Table 8.2.

We report statistical significance via one-way ANOVA, Tukey's honestly significant difference (HSD) test, and two-tailed t-test. For all four conditions, one-way ANOVA reports significant differences for all four constructs: behavior naturalness ($F = 25.7, p < 0.001$), behavior human-likeliness ($F = 29.6, p < 0.001$), synchrony ($F = 22.0, p < 0.001$), and engagement ($F = 25.7, p < 0.001$). We perform a post-hoc pairwise comparison analysis with Tukey's HSD which reveals the following. Tukey's HSD shows statistically significant differences between all pairs ($p < 0.001$) except between the pairs of (IL-LSTM, *ASAP*) and (*ASAP*, *HI²-ADAM*) for the constructs of behavior naturalness and human-likeliness ($p = 0.22$ with $p = 0.52$ and $p = 0.12$ with $p = 0.37$ respectively). For synchrony and engagement constructs, all pairs are found to be significantly different ($p < 0.001$) except between the pair of (*ASAP*, *HI²-ADAM*) ($p = 0.9$ and $p = 0.9$ respectively). We also test the statistical significance by performing post-hoc t-test between all possible pairs of compared animations (or conditions) for each construct. Significant differences between all pairs ($p < 0.001$) except between the pair of (*ASAP*, *HI²-ADAM*) ($p = 0.98$) for all four constructs.

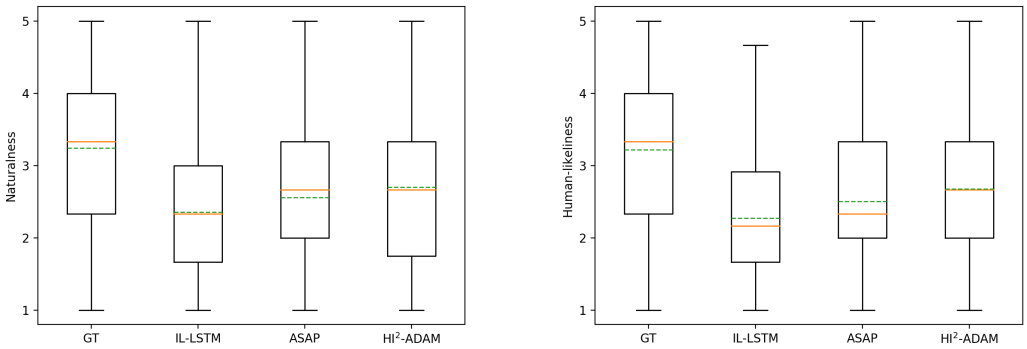


Figure 8.3 Distribution of behavior naturalness (left) and human-likeliness (right). Median represented by orange line and mean represented by green dashed line.

Looking at the results, we can note that the *SIA* visualization for the GT condition receives the highest values for all four constructs in terms of both median

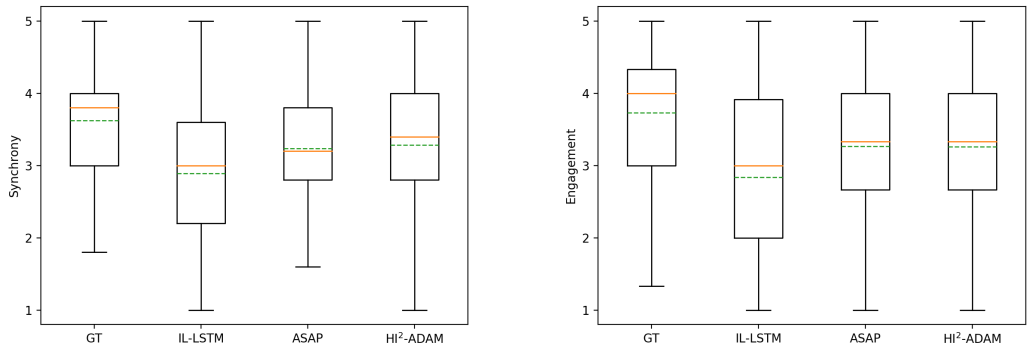


Figure 8.4 Distribution of synchrony (left) and engagement (right). Median represented by orange line and mean represented by green dashed line.

Methods	Naturalness	Human-likeliness	Synchrony	Engagement
GT	3.33/3.24	3.33/3.22	3.80/3.62	4.00/3.73
IL-LSTM	2.33/2.36	2.17/2.27	3.00/2.89	3.00/2.84
ASAP (ours)	2.67/2.56	2.33/2.51	3.20/3.23	3.33/3.26
HI²-ADAM (ours)	**2.67/2.70**	**2.67/2.68**	**3.40/3.28**	**3.33/3.26**

Table 8.2 Median/mean values of naturalness, human-likeliness, synchrony, and engagement.

and mean values: behavior naturalness (3.33, 3.24 respectively), behavior human-likeliness (3.33, 3.22 respectively), synchrony (3.80, 3.62 respectively), and engagement (4.00, 3.73 respectively).

The global trend of the increase in quality in the order of IL-LSTM, *ASAP*, and *HI²-ADAM* is noticeable across the 4 constructs of behavior naturalness, behavior human-likeliness, synchrony, and engagement. A clear difference can be observed between IL-LSTM and the two computational models of *ASAP* and *HI²-ADAM*. IL-LSTM reports the lowest median and mean values for all constructs: behavior naturalness (2.33, 2.36 respectively), behavior human-likeliness (2.17, 2.27 respectively), synchrony (3.00, 2.89 respectively), and engagement (3.00, 2.84 respectively). We assume that the difference in behavior naturalness and human-likeliness derives from the method of inferring the predictions. *ASAP* and *HI²-ADAM* both apply adaptive online prediction which generates *continuous* motions while IL-LSTM employs sliding window prediction. The higher perception of the constructs of naturalness and human-likeliness for the two computational models, *ASAP* and *HI²-ADAM*, may have been due to the generation of continuous behavior motions. For the synchrony and engagement constructs, the modeling of *reciprocal adaptation* (or *interpersonal relationship*) tends to lead to the rise in these constructs (Biancardi et al. [2021]) for *ASAP* and *HI²-ADAM* compared to IL-LSTM. We observe that such *reciprocal adaptation* modeling increases the synchrony and engagement between *SIA*s and their interlocutors. This observation validates our qualitative results and further demonstrates that the *behavioral appropriateness* (linked to the perception of behavior naturalness and human-likeliness) and *recip-*

rocal adaptation resemblance (linked to the perception of interpersonal dynamics of synchrony and engagement) are enhanced by the modeling of *behavior continuity* and *reciprocal adaptation* (i.e. *ASAP* and *HI²-ADAM*). We remark the same observation, as for *ASAP* (ref. Chapter 6), that such *reciprocal adaptation* modeling increases the synchrony and engagement between *SIA*s and their interlocutors.

Between *ASAP* and *HI²-ADAM*, we remark an augmentation of values for the constructs of naturalness (median: 2.67, mean: 2.56 and median: 2.67, mean: 2.70 respectively), human-likeliness (median: 2.33, mean: 2.51 and median: 2.67, mean: 2.68 respectively) and synchrony (median: 3.20, mean: 3.23 and median: 3.40, mean: 3.28 respectively). For the engagement, similar values are found for both *ASAP* (median: 3.33, mean: 3.26) and *HI²-ADAM* (median: 3.33, mean: 3.26). We note that the *HI²-ADAM* generates behaviors that elicit a higher perception of behavior naturalness and human-likeliness. We assume that it is due to the modeling of *intrapersonal relationship*, notably via the modeling of the *history* of each modality. This is inline with the previously reported objective ablation study, where the *behavior appropriateness*, in terms of behavior distribution similarity via KS test, is improved by the addition of the modality memory encoder E_m. Also, the deeper encoding of the *interpersonal dynamics*, via our interpersonal encoder E_{inter}, we notice a further improvement of the sync between the *SIA* and the human participant. We can check the same impact of adding the interpersonal encoder in the qualitative ablation study where the *reciprocal adaptation resemblance*, seen by the metrics of DTW and reciprocal adaptation measures (Sync and EL), is enhanced by the inclusion of the interpersonal encoder. While the statistical results do not show significant differences between *ASAP* and *HI²-ADAM*, the results of *HI²-ADAM* do not imply a fall in performance compared to *ASAP* but rather a rising tendency of the perceived human-likeliness and synchrony via the median values.

Our subjective evaluation results obtained through the user perception study are aligned with our objective evaluation. Our *HI²-ADAM* model outperforms the baseline models of LSTM and *ASAP* being the closest one to the GT, notably in terms of human-likeliness and synchrony. We validate that *HI²-ADAM* is capable of improving *interpersonal dynamics* of synchrony and engagement by capturing *interpersonal relationship* (via the interpersonal encoder E_{inter}). Also, the encoding of each *modality's history* (or memory) and the relationship between the *modality histories* (within the intrapersonal encoder E_{intra}) ameliorates the behavior naturalness and human-likeliness.

8.3 Contributions and Conclusion

8.3.1 Contributions

Our contributions are as follows:

- We propose *HI²-ADAM*, an approach to capture the *reciprocal adaptation* of *SIA* behaviors.

- We explicitly model the *intrapersonal* and *interpersonal relationships* by encoding prior emitted *multimodal* signals and their history via our *modality memory* encoders.

8.3.2 Conclusion

In this work, we propose a new approach to generate *adaptive SIA* behavior as both *speaker* and *listener* by encoding the *multimodality* and *intrapersonal* and *interpersonal relationships*. We conclude that *HI^2-ADAM* model outperforms state-of-the art approaches both quantitatively and qualitatively in producing *SIA* behavior for both *speaker* and *listener*, notably in terms of *reciprocal adaptation resemblance*, *behavior human-likeliness*, and *synchrony*.

The key points of this Chapter:

Addressing Research Questions

- *Interpersonal dynamics* of synchrony and engagement can be augmented by modeling the *interpersonal relationship* (or *reciprocal adaptation*).
- *Behavioral aspects* of naturalness and human-likeliness can be improved by capturing the *intrapersonal relationship* with the modelization of the relation between *modality histories*.

HI^2-ADAM Model

- The *reciprocal adaptation* was modeled via *HI^2-ADAM* model to generate *adaptive SIA* behavior as both *speaker* and *listener* by encoding the *multimodality* and *intrapersonal* (explicit modeling of *modality histories*) and *interpersonal relationships*.
- *HI^2-ADAM* produces *natural* and *engaging* behaviors that show remarkable performance in terms of *reciprocal adaptation resemblance* (and interlocutor *synchrony*) and behavior *human-likeliness*.

Publications

- *(Preprint)* - Jieyeon Woo, Mireille Fares, Catherine Pelachaud, and Catherine Achard. Amii: Adaptive multimodal inter-personal and intra-personal model for adapted behavior synthesis. *arXiv preprint arXiv:2305.11310*, 2023a

Chapter 9

Conclusion

Contents

In this thesis, we work on developing an adaptive Socially Interactive Agent with *reciprocal adaptation* capabilities acting as an interactive conversational partner that is *social*, *engaging*, and perceived as *natural* and *human-like*. To accomplish this challenging task, we started by analyzing the adaptation present in human-human interactions and proposed new *measures of reciprocal adaptation* (ref. Chapter 5). After investigating the presence and role of adaptation, using the investigation as a basis, we modeled the *reciprocal adaptation* to generate adaptive nonverbal *SIA* behaviors via two computational models of *ASAP* (ref. Chapter 6) and *HI*^{2}*-ADAM* (ref. Chapter 8). To evaluate the impact of *reciprocally adaptive SIA* behaviors in *real time*, a *real-time* system, *IAVA* system, of a *SIA* endowed with the adaptation capability was implemented and showcased for the applications of Cognitive Behavior Therapy and Social Skills Training (SST) (ref. Chapter 7).

This Chapter concludes the thesis. It starts by summarizing the contributions of this thesis. Then, the limitations of this work and the future directions of research are presented.

9.1 Summary of Contribution

This thesis contributes to the research communities of *SIA* and *multimodal* signal processing for generating *adaptive* nonverbal *SIA* behaviors by capturing *intrapersonal* and *interpersonal relationships* from *multimodal* signals. We discuss in detail the contributions made by this thesis.

Proposition of novel reciprocal adaptation measures

From the analysis of human-human interaction, we studied the adaptation present in conversations. The investigation served as a foundation to propose novel *measures of reciprocal adaptation*. The *reciprocal adaptation measures*, consisting of synchrony and entrainment loop measures, were employed to study the relationship between adaptation and the dimensions of engagement, warmth, and competence. The newly proposed measures showed their usefulness in assessing the human-agent interaction quality. They were used for the objective evaluations of our *IAVA* system (ref. Chapter 7) and *HI^2-ADAM* model (ref. Chapter 8).

Rendering adaptive SIA behaviors

The reciprocal adaptation capability, which is an important capacity innate to humans for interactive and engaging communications, is endowed to *SIA*s by modeling *multimodality*, *interpersonal* relationship, and/or *intrapersonal* relationship. We generate adaptive *SIA* behaviors via our *ASAP* model (ref. Chapter 6) and *HI^2-ADAM* model (ref. Chapter 8). The rendered *SIA* behavior is shown to outperform state-of-the-art techniques in generating *natural*, *human-like*, *in sync*, and *engaging* behavior. Through our *reciprocal adaptation measures* (ref. Chapter 5), we were also able to objectively validate that the predicted behaviors were indeed *reciprocally adaptive* and the usefulness of these measures in assessing the quality of human-agent interaction.

Development of a real-time interactive and adaptive SIA system

The ultimate goal when developing embodied agents, both *SIA*s and robots, is to deploy them in *real time* with the human end-user. The *real-time* operation is essential, especially for the adaptation in human-agent or human-robot interaction. We created an *interactive* and *adaptive SIA* system, our *IAVA* system (ref. Chapter 7), that assures the r*eal-time* aspect. By applying the *IAVA* system to the medical application of CBT, we verified the efficiency of *SIA*s with *reciprocal adaptation* capabilities in giving a positive impression to the users (being perceived as *natural*, *human-like*, *engaging*, *in sync*, and *building a rapport*) and in improving the CBT effect. Furthermore, to demonstrate the possibility of employing our adaptive *SIA* system in other applications, we have also tested our system for SST. We found that a similar impression of the agent is given to the users for both applications of SST and CBT despite their different nature of scenarios. Moreover, we collected a human-agent interaction database (*CBT-HAI DB*). The CBT interactions between

the *SIA* and the user were recorded and the database was made available to the research community (available after signing the EULA form).

9.2 Limitations and Future Work

This thesis tackles the challenge of creating adaptive *SIA*s that can function in *real time*. While the core challenges were addressed, there is still room for improvement. Here we present some limitations of our work and propositions to move forward.

Training database

The NoXi database (ref. Chapter 4) was used to train our adaptive *SIA* behavior generation models of *ASAP* (ref. Chapter 6) and *HI²-ADAM* (ref. Chapter 8). The database, made up of human-human interactions, has a total duration of only $7h22min$. The database is quite small for the gesture generation task. We addressed this problem of having such a small database via our modeling technique. Nevertheless, it will be helpful to have a bigger database to create more adaptive *SIA* behaviors. This could be done using the entire NoXi database of all three locations (French, English, and German with a total duration of $25h18min$). However, this risks generalizing the cultural aspects of the three countries, and the prosodic information has to be addressed differently as the spoken language is different. Another solution is to create a new corpus of dyadic interactions with a setting similar to the NoXi database.

Generation of adaptive SIA gestures

In this work, adaptive *SIA* behaviors, consisting of facial expressions and head-/gaze movements, were generated. In addition to these behaviors, the adaptive behavior generation can be extended to the *SIA*'s full body motion (notably for upper body gestures including hand gestures and torso movements) by training the models (*ASAP* and *HI²-ADAM*) with the full body data. Furthermore, the full body adaptive gestures can be visualized in real time by extending the *IAVA* system and by integrating generation models trained for the full body to it. In addition, the adaptive prosodic behavior can be modeled and endowed to the *SIA* such as laughter and voice prosody. The adaptive behavior generation models for such diverse behavior types can also be parameterized to simulate unique individuals with different social attitudes and personality traits.

Merging different levels of adaptation

This thesis mainly works on modeling the adaptation at the signal level. However, adaptation can be found at multiple levels from low (i.e. signal level) to high (i.e. context/intent level). The next step may be to mix both high-level and low-level adaptations and give this skill to the *SIA*. Also, the endowment of adaptive verbal and nonverbal behavior may allow the *SIA* to become a better interlocutor.

Cultural differences in adaptive behavior

Another possible direction of research is to investigate the adaptation between cultures. The culture may play an important role in the form, frequency/timing, and appropriateness of adaptation. It may be interesting to study this cultural difference. Also, we could test whether the adaptive behavior of one culture could be employed by the users of another culture and be perceived in the same way.

Trustworthiness of AI models

The use of AI models like our adaptive behavior generation models, *ASAP* and *HI^2-ADAM*, may pose ethical concerns and risks such as the distribution of harmful content, data privacy violations, sensitive information disclosure, amplification of existing bias, and lack of explainability and interpretability. To prevent such problems, developers of AI models should be cautious when creating such models by following the regulations and requirements of EU's Artificial Intelligence Act (AIA). Also, they should provide enough information about their system to their users and conduct thorough pretests. Adding explicability to AI models would help the interpretation of the AI's decision and increase the users' trustworthiness towards AI.

Part VI

Appendices

Chapter 10

Appendix A

Facial gestures (eyebrow, cheek, and mouth movements) can be characterized/annotated by facial landmarks or Action Units (AUs; Ekman and Friesen [1976]).

10.1 Face landmarks

Face landmarks, shown in Figure 10.1 are key attributes of a human face such as eyebrows, nose, mouth, and eye corners. They allow the identification of an individual by distinguishing from different faces. Such facial landmarks have been successfully used for various computer vision applications of face alignment, face swapping, and emotion detection.

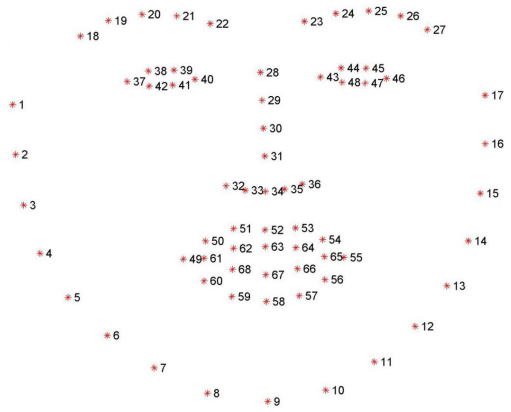


Figure 10.1 Illustration of facial landmarks (68 facial landmark coordinates). Image from pyimagesearch [2017].

10.2 Facial AUs

FACS (Facial Action Coding System), developed by Ekman and Friesen [1976], is a facial muscle scheme based on manual facial expression analysis. It is a technique to interpret facial expressions by dividing facial muscle movements into 46 AUs,

as illustrated in Figure 10.2. Each AU matches a certain facial muscle movement or expression. Facial gestures can be expressed by a single AU or a combination of AUs, as shown in Figure 10.3. Especially for the production of emotions, multiple AUs are combined to formulate complex expressions. AUs have been used in various studies notably for emotion studies.

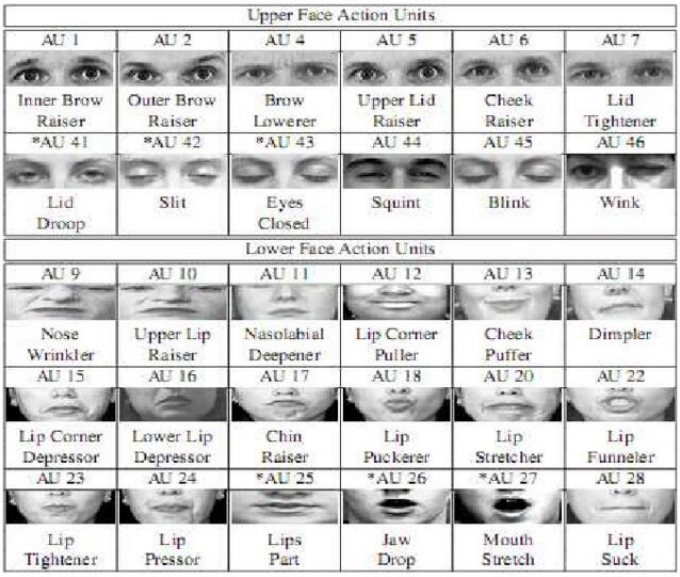


Figure 10.2 Illustration of facial action units (Li et al. [2005]).

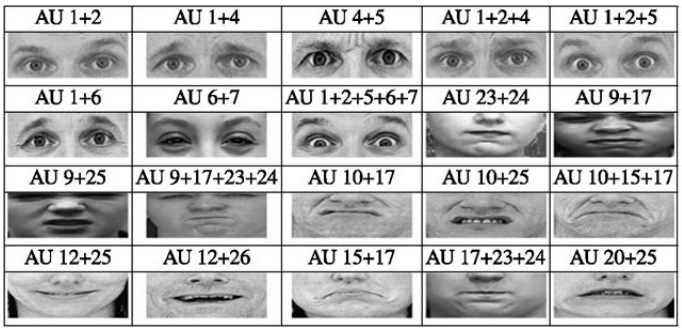


Figure 10.3 Illustration of different combinations of facial action units (Li et al. [2005]).

Chapter 11

Appendix B

We report here additional information regarding the *IAVA* system architecture (Woo et al. [2023c]) described in Chapter 7.

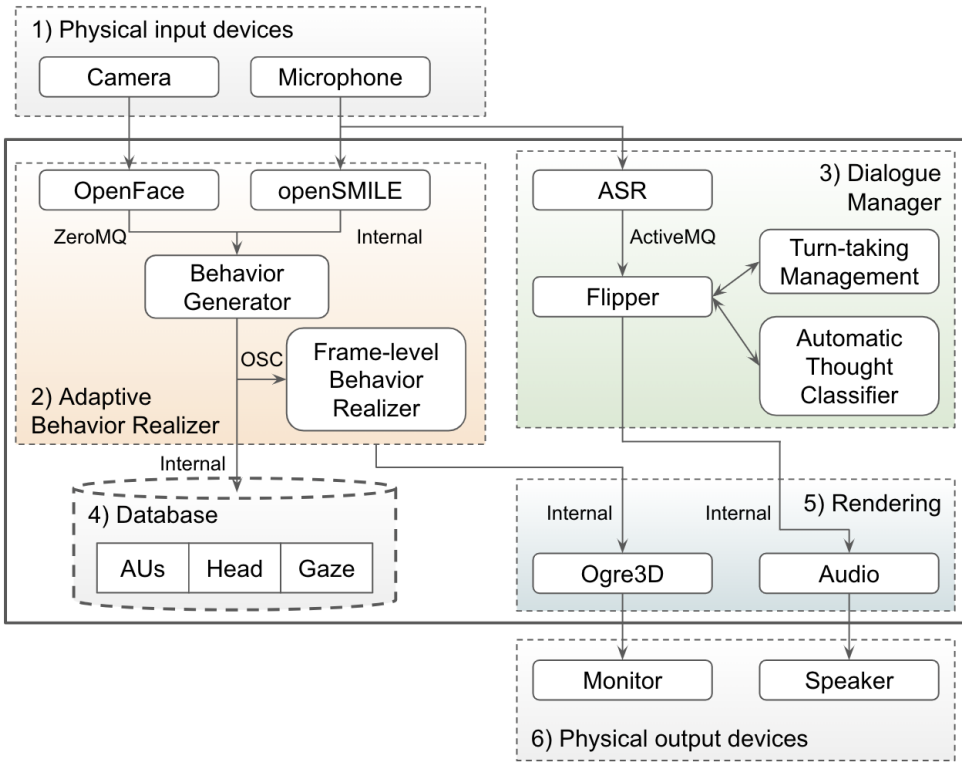


Figure 11.1 *IAVA* system architecture.

IAVA is composed of six parts, as illustrated in Figure 11.1, which are: 1) physical input devices; 2) Adaptive Behavior Realizer; 3) Dialogue Manager; 4) database; 5) rendering; and 6) physical output devices.

11.1 System Inputs and Outputs

Our system makes use of various physical devices as input and output, and communicates multiple signals via different communication protocols.

11.1.1 Physical devices

The system uses a 1080p RGB webcam to capture the user's face, a pin microphone to capture the user's speech, a speakerphone to render the agent's speech utterance, and a monitor to display the virtual agent (in a close-up of their face, head, and shoulders) as shown in Figure 11.2.

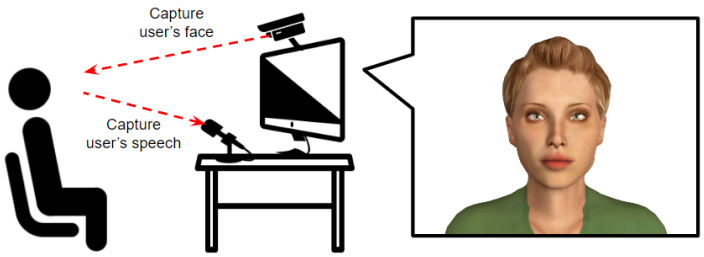


Figure 11.2 The system is equipped with a webcam to capture the user's face and a microphone to capture the user's speech. The virtual agent is displayed in front of the user.

11.1.2 Signals

The input and output signals communicated within the system are as follows.

Visual features: The visual features of the user are extracted in real time at $30fps$ by processing the webcam-rendered images of the user using OpenFace. To be more specific, the visual features of eye movements (around the x and y axes), head rotations (around the x, y, and z axes), 6 upper face AUs (which are *AU1*, *AU2*, *AU4*, *AU5*, *AU6*, and *AU7*) along with that of the smile (*AU12*) are passed to the model to generate the agent behavior.

Audio features: The audio features of both human user and agent are obtained separately in real time at *100 Hz* from the user's speech captured by the microphone via openSMILE. To detail, the fundamental frequency, loudness, voicing probability, and 13 MFCCs are fed to the model for the prediction.

Utterance text: The text of the user's utterance is acquired by ASR from the microphone captured user's speech. The text utterance is given as input to the Flipper engine to manage the dialogue.

Agent animation: The agent animation realized for each frame is visualized with Ogre3D and displayed on the monitor.

Agent speech: The selected agent's speech is generated via the Greta platform's Audio module, to transform the text selected by the Dialogue Manager into audio, and the audio is rendered with the speakerphone.

11.1.3 Communication protocols

The signals are passed between different toolkits and modules via communication protocols which are:

ZeroMQ: ZeroMQ [1] (Hintjens [2013]) is an asynchronous network messaging library that is used for distributed and concurrent systems. Messages such as binary data, serialized data, and simple strings can be sent without a dedicated message broker. In our system, it is used to transmit real-time OpenFace signals directly to the model.

ActiveMQ: ActiveMQ [2] (Snyder et al. [2011]) is an open-source message broker which can foster multi-client or multi-server communication. *IAVA* employs ActiveMQ messages to send the user's utterance recognized by the ASR to Flipper.

OSC: OSC (Open Sound Control) [3] (Wright [2005]) is a lightweight and flexible protocol for real-time message communication. The advantages of OSC are its possibility to receive signals from other computers and platforms, and its availability in multiple programming languages. Our system makes use of OSC to communicate between the computational model externally running in Python and the Ogre3D of the Greta platform operating in Java.

11.2 Adaptation Behavior Realizer

To generate real-time adaptive behavior, we implement the Adaptation Behavior Realizer (ABR) module. The ABR module consists of two main components which are the Behavior Generator module and the Frame-level Behavior Realizer module as seen in Figure 11.3.

11.2.1 Behavior Generator module

The Behavior Generator module integrates a pre-trained computational model, *ASAP* model (Woo et al. [2023d]), which generates the agent behavior that is reciprocally adaptative. The model takes the past 100 time-steps of both the human user's and the agent's behavior (visual and audio features) to predict the agent's visual behavior at the next time-step. The *ASAP* model learns interpersonal relationship from real human-human interactions (Cafaro et al. [2017]). It

[1] https://zeromq.org
[2] https://activemq.apache.org
[3] https://opensoundcontrol.org

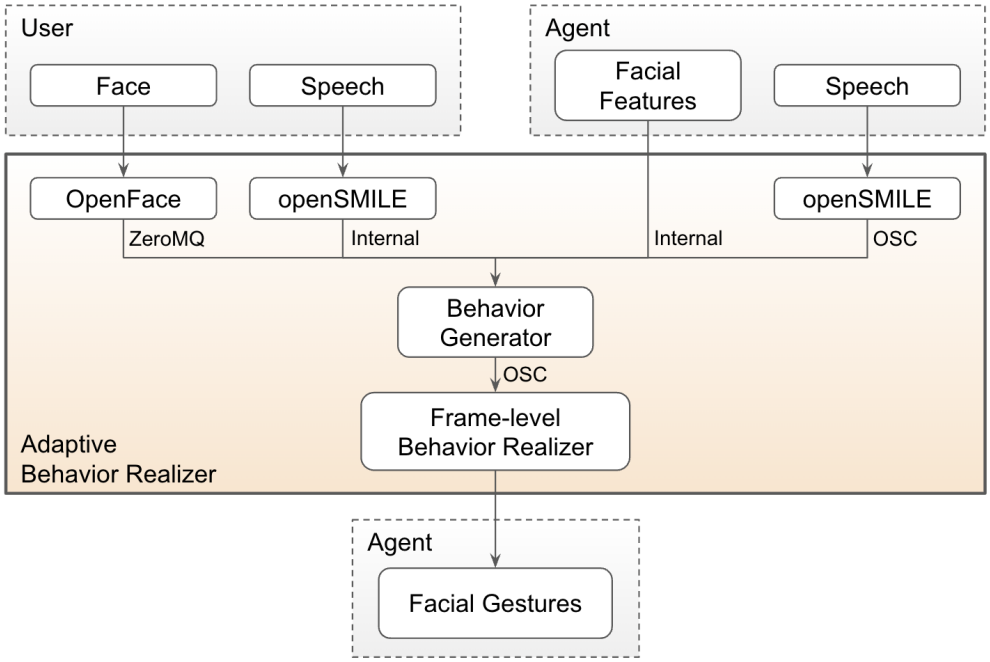


Figure 11.3 The Adaptation Behavior Realizer generates the agent's adaptive behavior and visualizes it at the frame-level. The agent's behavior is predicted with the Behavior Generator module via the *ASAP* model (Woo et al. [2023d]) which considers the face and speech signals from both the human user and agent of the past time-steps. The generation is then rendered for each frame at $25fps$ via the Frame-level Behavior Realizer module.

models the reciprocal adaptation capability and endows it to the agent from multimodal signals exchanged within a dyadic interaction with its data augmentation and self-attention pruning techniques. It generates the agent's adaptive behavior (outputting facial AUs and head/gaze movements) while assuring movement continuity via autoregressive adaptive online prediction for every frame (at each time-step) at $25fps$.

To obtain the agent's behavior at $25Hz$, the Behavior Generator module first extracts the features individually with different sampling rates as the following:

- User's audio features via openSMILE at $100Hz$ and communicated internally;
- User's visual features via OpenFace at $30Hz$ and communicated with ZeroMQ;
- Agent's audio features via openSMILE at $100Hz$ and communicated with OSC;
- Agent's visual features via the computational model at $25Hz$ and communicated internally.

We sync the different sampling rates to $25Hz$ (i.e. $25fps$) which is the computational model's sampling rate. The last 100 time-steps' signals are stocked and updated of all four feature categories with internal objects for the agent's behavior prediction of the next time-step. Each prediction, composed of the agent's facial

expression (*AU1*, *AU2*, *AU4*, *AU5*, *AU6*, *AU7*, and *AU12*), head rotations, and gaze, is sent via OSC to the Frame-level Behavior Realizer module to display the agent's behavior at the frame-level. After each prediction, the four feature categories of user and agent are saved into the database in a CSV format at the sampling rate of $25Hz$.

11.2.2 Frame-level Behavior Realizer

The Frame-level Behavior Realizer module receives the agent's behavior generated by the Behavior Generator module via OSC. The Greta platform's original Behavior Realizer module (Niewiadomski et al. [2009]) generates the agent's behavior by passing the user's raw input data through the Intent Planner module and the Behavior Planner module. It realizes the behavior in sequences that corresponds to the command sent by the Intent Planner. Our Frame-level Behavior Realizer module, which can be seen in Figure 11.3, differs from the original Behavior Realizer in the sense that it enables the generation of behaviors at the frame-level (at each time-step) which allows the virtual agent to continuously show smooth behavior throughout the whole interaction. Moreover, it produces the agent's behavior directly from the raw user input data. It is also possible to select the types of agent behavior that will be displayed via an interactive window. The types of agent behavior that can be activated are the following:

- Each upper face Action Unit (*AU1*, *AU2*, *AU4*, *AU5*, *AU6*, and *AU7*);
- Smile (*AU12*);
- Blink (*AU45*) which is automatically generated internally;
- Gaze (around the x and y axes);
- Head movement along each axis (x, y, and z);
- Mouth movement.

The *IAVA* system checks which agent behavior types are activated, at the beginning of the interaction, and displays them. For the ones that are deactivated, the agent will show the default behavior (value of 0 for the intensity of the AUs and 0 degrees for the head rotations and gaze angles). The selected combination of the agent's behavior is passed directly to the Ogre3D for rendering.

11.3 Dialogue Manager

The natural flow of the dialogue is managed by the Dialogue Manager. The dialogue is controlled by the Flipper engine which continuously communicates with the Turn-taking Management module, as illustrated in Figure 11.4, to choose the next conversational move. For the application of CBT, the Automatic Thought Classifier module is integrated into the Dialogue Manager. The process is as follows. Flipper first receives via ActiveMQ the utterance text of the user's response from the ASR. For each new utterance, it checks whether the utterance corresponds to

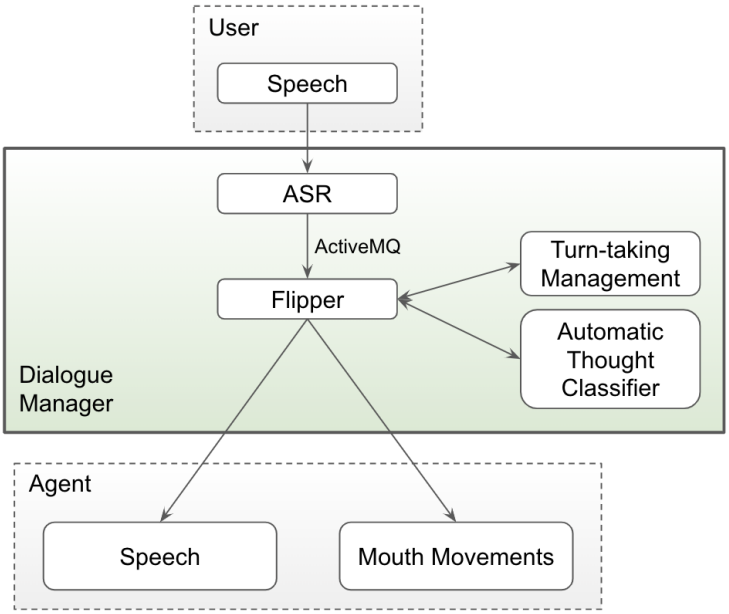


Figure 11.4 The Dialogue Manager manages the conversation dialogue. It selects the next conversational move while assuring the natural flow of the interaction by constantly communicating with the Turn-taking Management module. For the CBT application, the Automatic Thought Classifier module was integrated into the Dialogue Manager

an automatic thought via the Automatic Thought Classifier module and directs the conversational flow with the Turn-taking Management module. The modules are further explained below. The communicative intentions selected by Flipper are then instantiated into mouth movements which are combined and synchronized with the agent's speech via the Greta platform's standard treatment of passing by the Greta platform's original modules of Behavior Planner, Behavior Realizer, and Speech Synthesizer. The produced agent's mouth movements and speech are each sent to the Orgre3D and Audio module for rendering, as shown in Figure 11.1. This process is repeated for each user's utterance throughout the interaction.

11.3.1 Turn-taking Management module

Turn-taking is managed with the Turn-taking Management module to assure a smooth and natural flow of the conversation. The module keeps track of the speaking state of the agent and that of the human user. It handles conversational turn-taking by looking at both speaking states. By observing these two states, the agent interprets whether the user has finished answering and is giving their speaking turn (to address single responses made up of several utterances linked with pauses) and whether the user is reacting with backchannels (i.e. not aiming at taking the speaking turn), and thus decide when to take the speaking turn. After the agent decides to take the turn, it proceeds with its next conversational move.

11.3.2 Automatic Thought Classifier module

For CBT interaction, to proceed with the CBT scenario proposed in Shidara et al. [2022], a semantic analysis of the user's utterance needs to be done to identify whether the user has answered with an automatic thought or not. The structural content of the dialogue is processed by the Automatic Thought Classifier module. The module integrates the classifier model presented in Shidara et al. [2022] using the classifier algorithm of Support Vector Machine (SVM; linear kernel) with the French word embeddings from Bidirectional Encoder Representations from Transformers (BERT; Devlin et al. [2018]), which is a pre-trained language model for word representations. As in Shidara et al. [2022], the raw text is tokenized and a part-of-speech tag is associated with each token. All input sentences are covered with [CLS] and [SEP] tokens, which are placed at the beginning and the ending respectively, and are fed to BERT with a hidden vector of 768 dimensions. These tags are used as the inputs of the classifier model. The model identifies automatic thoughts by performing binary classification on the user's utterance. Depending on whether the user's response is an automatic thought or not, the next agent's utterance is decided.

11.4 Animation Rendering

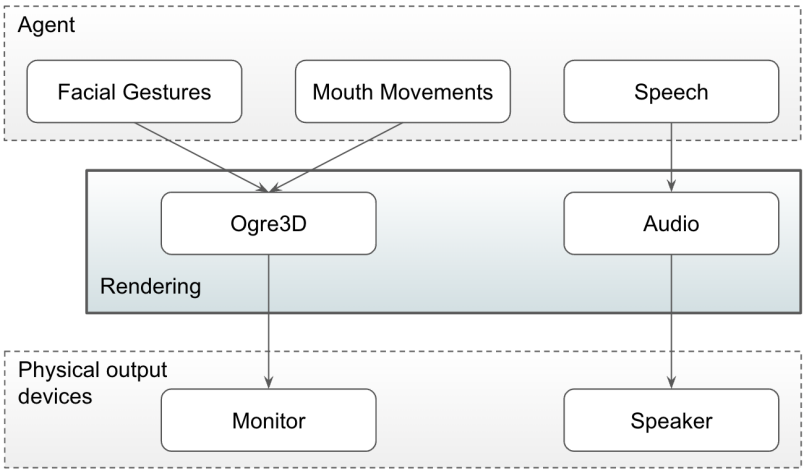


Figure 11.5 The Animation Rendering module displays the generated agent's behaviors, which are the agent's facial gestures obtained by the Adaptation Behavior Realizer and the agent's mouth movements sent by the Dialogue Manager, and renders the agent's speech produced by the Dialogue Manager.

The final animation of the generated agent's behavior, which consists of the agent's facial gestures obtained at the frame-level by the Adaptation Behavior Realizer and the agent's mouth movements and speech produced by the Dialogue Manager is rendered by the Animation Rendering module. The agent's facial gestures and mouth movements, visualized together via Ogre3D, and the agent's ut-

terance, produced by Greta platform's Audio module, are each passed to their corresponding physical output devices (monitor and speaker respectively).

11.5 System Performance and Specifications

The *IAVA* system works in real time, executing a single system loop every $0.04s$. The single system loop consists of:

- perception of $0.03s$ with OpenFace at $30fps$ and openSMILE at $100Hz$;
- adaptive behavior generation of $< 0.01s$ via the *ASAP* model;
- communication and visualization of $< 0.01s$.

All signals within the system are synced without any delay for it to function in $25Hz$, and thus generate and display the agent's behavior every $0.04s$.

For the functioning of the system, a space requirement of approximately $7GB$ is needed which consists of: $2GB$ for platform visualization, $2GB$ for OpenFace and openSMILE, and $3GB$ for execution and data saving.

In addition to the memory space requirement, hardware specifications must be met which are two computers with $2.4GHz$ Intel Core i9 mounted with NVIDIA Quadro RTX 4000 and $64GB$ RAM.

Chapter 12

Appendix C

We present here the interpretation of correlation graphs (Figures 7.7, 7.8, and 7.9) presented in Chapter 7.

The graphs were obtained via Jamovi, an open-source statistical software [1] (Şahin and Aybek [2019]).

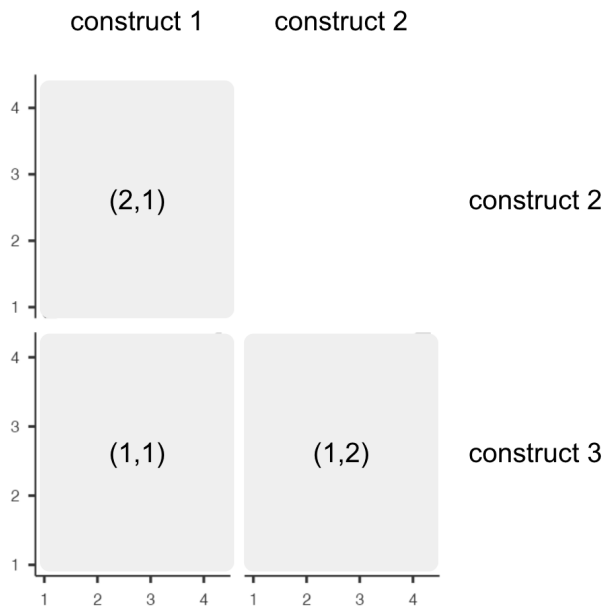


Figure 12.1 Example correlation graph for the interpretation explication.

For the explication of the graph interpretation, we use the example correlation graph shown in Figure 12.1. The example graph shows 3 relations between the constructs 1, 2, and 3. Each subgraph shows the correlation between:

- Subgraph (2,1): constructs 1 & 2;
- Subgraph (1,1): constructs 1 & 3;
- Subgraph (1,2): constructs 2 & 3.

[1] https://www.jamovi.org/

Bibliography

Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019.

Chaitanya Ahuja, Shugao Ma, Louis-Philippe Morency, and Yaser Sheikh. To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations. In *2019 International Conference on Multimodal Interaction*, pages 74–84, 2019.

Samer Al Moubayed, Jonas Beskow, and Gabriel Skantze. The furhat social companion talking head. In *INTERSPEECH*, pages 747–749, 2013.

Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. In *Computer Graphics Forum*, volume 39, pages 487–496. Wiley Online Library, 2020.

Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5223–5232, 2020.

Sadegh Aliakbarian, Fatemeh Saleh, Lars Petersson, Stephen Gould, and Mathieu Salzmann. Contextually plausible and diverse 3d human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11333–11342, 2021.

Keith Anderson, Elisabeth André, Tobias Baur, Sara Bernardini, Mathieu Chollet, Evi Chryssafidou, Ionut Damian, Cathy Ennis, Arjan Egges, Patrick Gebhard, et al. The tardis framework: intelligent virtual agents for social coaching in job interviews. In *Advances in Computer Entertainment: 10th International Conference, ACE 2013, Boekelo, The Netherlands, November 12-15, 2013. Proceedings 10*, pages 476–491. Springer, 2013.

Rita B Ardito and Daniela Rabellino. Therapeutic alliance and outcome of psychotherapy: historical excursus, measurements, and prospects for research. *Frontiers in psychology*, 2:270, 2011.

Michael Argyle. *Bodily communication*. Routledge, 2013.

Ted P Asay and Michael J Lambert. The empirical case for the common factors in therapy: Quantitative findings. 1999.

Kathleen T Ashenfelter, Steven M Boker, Jennifer R Waddell, and Nikolay Vitanov. Spatiotemporal symmetry and multifractal structure of head movements during dyadic conversation. *Journal of Experimental Psychology: Human Perception and Performance*, 35(4):1072, 2009.

Team Audacity. Audacity. *The Name Audacity (R) Is a Registered Trademark of Dominic Mazzoni Retrieved from http://audacity. sourceforge. net*, 2017.

Jeremy N Bailenson and Nick Yee. Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological science*, 16 (10):814–819, 2005.

Hanieh Bakhshayesh, Sean P Fitzgibbon, Azin S Janani, Tyler S Grummett, and Kenneth J Pope. Detecting synchrony in eeg: A comparative study of functional connectivity measures. *Computers in biology and medicine*, 105:1–15, 2019.

Gene Ball, Jack Breese, et al. Emotion and personality in a conversational agent. *Embodied conversational agents*, 189, 2000.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.

George Bebis and Michael Georgiopoulos. Feed-forward neural networks. *Ieee Potentials*, 13(4):27–31, 1994.

Aaron T Beck. *Cognitive therapy of depression*. Guilford press, 1979.

Judith S Beck. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications, 2020.

Alan S Bellack, Kim T Mueser, Susan Gingerich, and Julie Agresta. *Social skills training for schizophrenia: A step-by-step guide*. Guilford Publications, 2013.

Štefan Beňuš, Agustín Gravano, and Julia Hirschberg. Pragmatic aspects of temporal accommodation in turn-taking. *Journal of Pragmatics*, 43(12):3001–3027, 2011.

Frank J Bernieri and Robert Rosenthal. Interpersonal coordination: Behavior matching and interactional synchrony. 1991.

Frank J Bernieri, J Steven Reznick, and Robert Rosenthal. Synchrony, pseudosynchrony, and dissynchrony: measuring the entrainment process in mother-infant interactions. *Journal of personality and social psychology*, 54(2):243, 1988.

Michael J Bernstein, Donald F Sacco, Christina M Brown, Steven G Young, and Heather M Claypool. A preference for genuine smiles following social exclusion. *Journal of experimental social psychology*, 46(1):196–199, 2010.

Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents** this work has been supported in part by aro grants w911nf1910069 and w911nf1910315, and intel. code and additional materials available at: https://gamma. umd. edu/t2g. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 1–10. IEEE, 2021.

Beatrice Biancardi, Angelo Cafaro, and Catherine Pelachaud. Analyzing first impressions of warmth and competence from observable nonverbal cues in expert-novice interactions. In *Proceedings of the 19th acm international conference on multimodal interaction*, pages 341–349, 2017.

Beatrice Biancardi, Soumia Dermouche, and Catherine Pelachaud. Adaptation mechanisms in human–agent interaction: Effects on user's impressions and engagement. *Frontiers in Computer Science*, 3:696682, 2021.

Timothy Bickmore. Health-related applications of socially interactive agents. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application*, pages 403–436. 2022.

Steven M Boker, Jennifer L Rotondo, Minquan Xu, and Kadijah King. Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological methods*, 7(3):338, 2002.

Guy A Boy. *The handbook of human-machine interaction: a human-centered design approach*. CRC Press, 2017.

Jeffrey M Bradshaw, Paul Feltovich, and Matthew Johnson. Human-agent interaction. *Handbook of human-machine interaction*, pages 283–302, 2017.

Judee K Burgoon, Lesa A Stern, and Leesa Dillman. *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press, 1995.

Judee K Burgoon, Laura K Guerrero, and Valerie Manusov. Nonverbal signals. *The SAGE handbook of interpersonal communication*, pages 239–280, 2011.

Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth Andre, and Michel Valstar. The noxi database: multimodal recordings of mediated novice-expert interactions. pages 350–359, 11 2017. doi: 10.1145/3136755.3136780.

Yujun Cai, Yiwei Wang, Yiheng Zhu, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Chuanxia Zheng, Sijie Yan, Henghui Ding, et al. A unified 3d human motion synthesis model via conditional variational auto-encoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11645–11655, 2021.

Nick Campbell. Multimodal processing of discourse information; the effect of synchrony. In *2008 Second International Symposium on Universal Communication*, pages 12–15. IEEE, 2008.

Joseph N Cappella. Mutual adaptation and relativity of measurement. *Studying interpersonal interaction*, 1:103–117, 1991.

Joseph N Cappella. Behavioral and judged coordination in adult informal social interactions: Vocal and kinesic indicators. *Journal of personality and social psychology*, 72(1):119, 1997.

Justine Cassell. Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine*, 22(4):67–67, 2001.

Justine Cassell, Tim Bickmore, Lee Campbell, Hannes Vilhjalmsson, and Hao Yan. Human conversation as a system framework: Designing embodied conversational agents. *Embodied conversational agents*, pages 29–63, 2000.

Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. Beat: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 477–486, 2001.

William I. Chang and Eugene L. Lawler. Sublinear approximate string matching and biological applications. *Algorithmica*, 12(4):327–344, 1994.

Tanya L Chartrand and John A Bargh. The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893, 1999.

Tanya L Chartrand and Jessica L Lakin. The antecedents and consequences of human behavioral mimicry. *Annual review of psychology*, 64:285–308, 2013.

Jonas Chatel-Goldman, Marco Congedo, Christian Jutten, and Jean-Luc Schwartz. Touch increases autonomic coupling between romantic partners. *Frontiers in behavioral neuroscience*, 8:95, 2014.

Hang Chu, Daiqing Li, and Sanja Fidler. A face-to-face neural conversation model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7113–7121, 2018.

William S Condon and William D Ogston. Sound film analysis of normal and pathological behavior patterns. *Journal of nervous and mental disease*, 1966.

William S Condon and William D Ogston. A segmentation of behavior. *Journal of psychiatric research*, 5(3):221–235, 1967.

William S Condon and Louis W Sander. Neonate movement is synchronized with adult speech: Interactional participation and language acquisition. *Science*, 183 (4120):99–101, 1974.

Amy JC Cuddy, Peter Glick, and Anna Beninger. The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in organizational behavior*, 31:73–98, 2011.

Kerstin Dautenhahn. Embodiment and interaction in socially intelligent life-like agents. In *International Workshop on Computation for Metaphors, Analogy, and Agents*, pages 102–141. Springer, 1998.

Emilie Delaherche and Mohamed Chetouani. Multimodal coordination: exploring relevant features and measures. In *Proceedings of the 2nd international workshop on Social signal processing*, pages 47–52, 2010.

Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365, 2012.

Soumia Dermouche and Catherine Pelachaud. Engagement modeling in dyadic interaction. In *2019 international conference on multimodal interaction*, pages 440–445, 2019a.

Soumia Dermouche and Catherine Pelachaud. Generative model of agent's behaviors in human-agent interaction. In *2019 International Conference on Multimodal Interaction*, pages 375–384, 2019b.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Chuang Ding, Lei Xie, and Pengcheng Zhu. Head motion synthesis from speech using deep neural networks. *Multimedia Tools and Applications*, 74(22):9871–9888, 2015.

Zijian Ding, Jiawen Kang, Tinky Oi Ting Ho, Ka Ho Wong, Helene H Fung, Helen Meng, and Xiaojuan Ma. Talktive: a conversational agent using backchannels to engage older adults in neurocognitive disorders screening. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

Alan Dix. *Human-computer interaction*. Pearson Education, 2003.

Sidney S D'Mello, Patrick Chipman, and Art Graesser. Posture as a predictor of learner's affective engagement. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, 2007.

Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976.

Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

Paul Ekman and Wallace V Friesen. Felt, false, and miserable smiles. *Journal of nonverbal behavior*, 6:238–252, 1982.

Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

Mireille Fares, Catherine Pelachaud, and Nicolas Obin. Transformer network for semantically-aware and speech-driven upper-face generation. In *EUSIPCO*, 2022.

Mireille Fares, Catherine Pelachaud, and Nicolas Obin. Zero-shot style transfer for multimodal data-driven gesture synthesis. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–4. IEEE, 2023.

Will Feng, Anitha Kannan, Georgia Gkioxari, and C Lawrence Zitnick. Learn2smile: Learning non-verbal interaction through observation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4131–4138. IEEE, 2017.

Ylva Ferstl, Michael Neff, and Rachel McDonnell. Multi-objective adversarial gesture generation. In *Motion, Interaction and Games*, pages 1–10. 2019.

Ylva Ferstl, Sean Thomas, Cédric Guiard, Cathy Ennis, and Rachel McDonnell. Human or robot? investigating voice, appearance and gesture motion realism of conversational social agents. In *Proceedings of the 21st ACM international conference on intelligent virtual agents*, pages 76–83, 2021.

Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Andrea Bönsch, and Willem-Paul Brinkman. The 19 unifying questionnaire constructs of artificial social agents: An iva community analysis. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2020.

Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, and Willem-Paul Brinkman. Questionnaire items for evaluating artificial social agents-expert generated, content validated and reliability analysed. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 84–86, 2021.

Paul M Fitts. Human engineering for an effective air-navigation and traffic-control system. 1951.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785, 2017.

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

Joseph Grafsgaard, Nicholas Duran, Ashley Randall, Chun Tao, and Sidney D'Mello. Generative multimodal models of nonverbal synchrony in close relationships. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 195–202. IEEE, 2018.

Jonathan Gratch and Gale Lucas. Rapport between humans and socially interactive agents. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*, pages 433–462. 2021.

Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18 (5):602–610, 2005. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2

005.06.042. URL https://www.sciencedirect.com/science/article/pii/S0893608005001206. IJCNN 2005.

Dennis Greenberger and Christine A Padesky. *Mind over mood: Change how you feel by changing the way you think*. Guilford Publications, 2015.

David Greenwood, Stephen Laycock, and Iain Matthews. Predicting head pose from speech with a conditional variational autoencoder. ISCA, 2017.

Aman Gupta, Finn L Strivens, Benjamin Tag, Kai Kunze, and Jamie A Ward. Blink as you sync: Uncovering eye and nod synchrony in conversation using wearable sensing. In *Proceedings of the 23rd International Symposium on Wearable Computers*, pages 66–71, 2019.

Amy G Halberstadt. Family expressiveness styles and nonverbal communication skills. *Journal of Nonverbal Behavior*, 8:14–26, 1983.

Joanna Hale, Jamie A Ward, Francesco Buccheri, Dominic Oliver, and Antonia F de C Hamilton. Are you on my wavelength? interpersonal coordination in dyadic conversations. *Journal of nonverbal behavior*, 44(1):63–83, 2020.

Jennifer Hamet Bagnou, Elise Prigent, Jean-Claude Martin, Jieyeon Woo, Liu Yang, Catherine Achard, Catherine Pelachaud, and Céline Clavel. A framework for the assessment and training of collaborative problem-solving social skills. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pages 381–384, 2021.

Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. Evaluation of speech-to-gesture generation using bi-directional lstm network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 79–86, 2018.

Alexander Heimerl, Tobias Baur, Florian Lingenfelser, Johannes Wagner, and Elisabeth André. Nova-a tool for explainable cooperative machine learning. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 109–115. IEEE, 2019.

Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020.

Matthew J Hertenstein, Rachel Holmes, Margaret McCullough, and Dacher Keltner. The communication of emotion via touch. *Emotion*, 9(4):566, 2009.

Ursula Hess and Patrick Bourgeois. You smile–i smile: Emotion expression in social interaction. *Biological psychology*, 84(3):514–520, 2010.

Ursula Hess, Martin G Beaupré, Nicole Cheung, et al. Who to whom and why–cultural differences and similarities in the function of smiles. *An empirical reflection on the smile*, 4:187, 2002.

Ursula Hess, Stephanie Houde, and Agneta Fischer. Do we mimic what we see or what we know. pages 94–107, 2014.

Pieter Hintjens. *ZeroMQ: messaging for many applications*. " O'Reilly Media, Inc.", 2013.

Michael J Hove and Jane L Risen. It's all in the timing: Interpersonal synchrony increases affiliation. *Social cognition*, 27(6):949–960, 2009.

Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. Learning backchannel prediction model from parasocial consensus sampling: a subjective evaluation. In *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*, pages 159–172. Springer, 2010.

Becky Inkster, Shubhankar Sarda, Vinod Subramanian, et al. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6 (11):e12106, 2018.

Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2020.

Aidan Jones and Ginevra Castellano. Adaptive robotic tutors that support self-regulated learning: A longer-term investigation with primary school children. *International Journal of Social Robotics*, 10:357–370, 2018.

Stanley E Jones and A Elaine Yarbrough. A naturalistic study of the meanings of touch. *Communications Monographs*, 52(1):19–56, 1985.

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017.

Ronald C Kessler, Gavin Andrews, Lisa J Colpe, Eva Hiripi, Daniel K Mroczek, S-LT Normand, Ellen E Walters, and Alan M Zaslavsky. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological medicine*, 32(6):959–976, 2002.

Alaa Khamis, Jun Meng, Jin Wang, Ahmad Taher Azar, Edson Prestes, Árpád Takács, Imre J Rudas, and Tamás Haidegger. Robotics and intelligent systems against a pandemic. *Acta Polytechnica Hungarica*, 18(5):13–35, 2021.

Kyoung-jae Kim. Financial time series forecasting using support vector machines. *Neurocomputing*, 55(1-2):307–319, 2003.

Everlyne Kimani, Timothy Bickmore, Ha Trinh, and Paola Pedrelli. You'll be great: Virtual agent-based cognitive restructuring to reduce public speaking anxiety. In *2019 8th international conference on affective computing and intelligent interaction (ACII)*, pages 641–647. IEEE, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Sotaro Kita. Cross-cultural variation of speech-accompanying gesture: A review. *Language and cognitive processes*, 24(2):145–167, 2009.

Mark L Knapp, Judith A Hall, and Terrence G Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013.

Kyveli Kompatsiari, Vadim Tikhanoff, Francesca Ciardo, Giorgio Metta, and Agnieszka Wykowska. The importance of mutual gaze in human-robot interaction. In *Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9*, pages 443–452. Springer, 2017.

Sander L Koole and Wolfgang Tschacher. Synchrony in psychotherapy: A review and an integrative framework for the therapeutic alliance. *Frontiers in psychology*, 7:862, 2016.

Neeraj Kumar and Govind Kumar Jha. A time series ann approach for weather forecasting. *Int J Control Theory Comput Model (IJCTCM)*, 3(1):19–25, 2013.

Sing Lau. The effect of smiling on person perception. *The Journal of social psychology*, 117(1):63–67, 1982.

N Pontus Leander, Tanya L Chartrand, and John A Bargh. You give me the chills: Embodied reactions to inappropriate amounts of behavioral mimicry. *Psychological science*, 23(7):772–779, 2012.

Aouragh Si Lhoussain, GUEDDAH Hicham, and YOUSFI Abdellah. Adaptating the levenshtein distance to contextual spelling correction. *International Journal of Computer Science and Applications*, 12(1):127–133, 2015.

Bo Li, Tara N Sainath, Khe Chai Sim, Michiel Bacchiani, Eugene Weinstein, Patrick Nguyen, Zhifeng Chen, Yanghui Wu, and Kanishka Rao. Multi-dialect speech recognition with a single sequence-to-sequence model. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4749–4753. IEEE, 2018.

Jamy Li. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, 77:23–37, 2015.

Stan Z Li, Anil K Jain, Ying-Li Tian, Takeo Kanade, and Jeffrey F Cohn. Facial expression analysis. *Handbook of face recognition*, pages 247–275, 2005.

Marco Lippi, Matteo Bertini, and Paolo Frasconi. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):871–882, 2013.

Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Transactions on Management Information Systems (TMIS)*, 4(4):1–28, 2013.

Beth Logan. Mel frequency cepstral coefficients for music modeling. In *In International Symposium on Music Information Retrieval*. Citeseer, 2000.

Max M Louwerse, Rick Dale, Ellen G Bard, and Patrick Jeuniaux. Behavior matching in multimodal communication is synchronized. *Cognitive science*, 36(8): 1404–1426, 2012.

Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100, 2014.

Birgit Lugrin, Catherine Pelachaud, and David Traum. *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*. Morgan & Claypool, 2021.

Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13309–13318, 2021.

Soroosh Mariooryad and Carlos Busso. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. *IEEE Transactions on Affective Computing*, 6(2):97–108, 2014.

Malia F Mason, Elizabeth P Tatkow, and C Neil Macrae. The look of love: Gaze shifts and person perception. *Psychological science*, 16(3):236–239, 2005.

Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.

Mohsen Mohammadi, Faraz Talebpour, Esmaeil Safaee, Noradin Ghadimi, and Oveis Abedinia. Small-scale building load forecast based on hybrid forecast engine. *Neural Processing Letters*, 48(1):329–351, 2018.

Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20(1):70–84, 2010.

Yasmin Moslem, Rejwanul Haque, and Andy Way. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*, 2023.

Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

Gonzalo Navarro. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88, 2001.

Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. *arXiv preprint arXiv:2204.08451*, 2022.

Paula M Niedenthal, Martial Mermillod, Marcus Maringer, and Ursula Hess. The simulation of smiles (sims) model: Embodied simulation and the meaning of facial expression. 2010.

Radoslaw Niewiadomski, Elisabetta Bevacqua, Maurizio Mancini, and Catherine Pelachaud. Greta: an interactive expressive eca system. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 1399–1400, 2009.

Ryota Nishimura, Norihide Kitaoka, and Seiichi Nakagawa. A spoken dialog system for chat-like conversations considering response timing. In *Text, Speech and Dialogue: 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 3-7, 2007. Proceedings 10*, pages 599–606. Springer, 2007.

John C Norcross and Michael J Lambert. Psychotherapy relationships that work iii. *Psychotherapy*, 55(4):303, 2018.

Magalie Ochs and Catherine Pelachaud. Socially aware virtual characters: The social signal of smiles. *IEEE Signal Processing Magazine*, 30(2):128–132, 2013.

Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. Engagement in human-agent interaction: An overview. *Frontiers in Robotics and AI*, 7:92, 2020.

Teresa K O'Leary, Elizabeth Stowell, Everlyne Kimani, Dhaval Parmar, Stefan Olafsson, Jessica Hoffman, Andrea G Parker, Michael K Paasche-Orlow, and Timothy Bickmore. Community-based cultural tailoring of virtual agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2020.

Olivier Oullier, Gonzalo C De Guzman, Kelly J Jantzen, Julien Lagarde, and JA Scott Kelso. Social coordination dynamics: Measuring human bonding. *Social neuroscience*, 3(2):178–192, 2008.

Alfonso Palmer, Juan Jose Montano, and Albert Sesé. Designing an artificial neural network for forecasting tourism time series. *Tourism management*, 27(5):781–790, 2006.

Amit Kumar Pandey and Rodolphe Gelin. A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robotics & Automation Magazine*, 25(3):40–48, 2018.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*, 2019.

Dhaval Parmar, Stefan Olafsson, Dina Utami, Prasanth Murali, and Timothy Bickmore. Designing empathic virtual agents: manipulating animation, voice, rendering, and empathy to create persuasive agents. *Autonomous agents and multi-agent systems*, 36(1):17, 2022.

Florian Pecune, Angelo Cafaro, Magalie Ochs, and Catherine Pelachaud. Evaluating social attitudes of a virtual tutor. In *Intelligent Virtual Agents: 16th International Conference, IVA 2016, Los Angeles, CA, USA, September 20–23, 2016, Proceedings 16*, pages 245–255. Springer, 2016.

Marc D Pell. Prosody–face interactions in emotional processing as revealed by the facial affect decision task. *Journal of Nonverbal Behavior*, 29:193–215, 2005.

Andre Pereira, Rui Prada, and Ana Paiva. Improving social presence in human-agent interaction. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1449–1458, 2014.

Carlos Pereira Santos, Joey Relouw, Kevin Hutchinson-Lhuissier, Alexander van Buggenum, Agathe Boudry, Annemarie Fransen, Myrthe van der Ven, and Igor Mayer. Embodied agents for obstetric simulation training. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 515–527, 2023.

Pierre Philip, Lucile Dupuy, Marc Auriacombe, Fushia Serre, Etienne de Sevin, Alain Sauteraud, and Jean-Arthur Micoulaud-Franchi. Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients. *NPJ digital medicine*, 3(1):2, 2020.

Martin J Pickering and Simon Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190, 2004.

Ken Prepin and Catherine Pelachaud. Basics of intersubjectivity dynamics: Model of synchrony emergence when dialogue partners understand each other. In *International Conference on Agents and Artificial Intelligence*, pages 302–318. Springer, 2011.

Ken Prepin, Magalie Ochs, and Catherine Pelachaud. Beyond backchannels: coconstruction of dyadic stancce by reciprocal reinforcement of smiles between virtual agents. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35, 2013.

pyimagesearch. Visualizing the 68 facial landmark coordinates from the ibug 300-w dataset, 2017. URL https://pyimagesearch.com/wp-content/uploads/2017/04/facial_landmarks_68markup.jpg. [Online; accessed September 15, 2023].

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

Stéphane Raffard, Robin N Salesse, Catherine Bortolon, Benoit G Bardy, José Henriques, Ludovic Marin, Didier Stricker, and Delphine Capdevielle. Using mimicry of body movements by a virtual agent to increase synchronization behavior and rapport in individuals with schizophrenia. *Scientific reports*, 8(1):17356, 2018.

Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrusaitis, and Roland Goecke. Extending long short-term memory for multi-view structured learning. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 338–353. Springer, 2016.

Fabian Ramseyer and Wolfgang Tschacher. Synchrony in dyadic psychotherapy sessions. *Simultaneity: Temporal structures and observer perspectives*, pages 329–347, 2008.

Fabian Ramseyer and Wolfgang Tschacher. Nonverbal synchrony of head-and body-movement in psychotherapy: different signals have different associations with outcome. *Frontiers in psychology*, 5:979, 2014.

Stéphane Rauzy, Mary Amoyal, and Béatrice Priego-Valverde. A measure of the smiling synchrony in the conversational face-to-face interaction corpus pacocheese. In *Workshop SmiLa, Language Resources and Evaluation Conference, LREC 2022*, 2022.

S Zahra Razavi, Lenhart K Schubert, Kimberly van Orden, Mohammad Rafayet Ali, Benjamin Kane, and Ehsan Hoque. Discourse behavior of older adults interacting with a dialogue agent competent in multiple topics. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 12(2):1–21, 2022.

Dennis Reidsma, Anton Nijholt, Wolfgang Tschacher, and Fabian Ramseyer. Measuring multimodal synchrony for human-computer interaction. In *2010 international conference on cyberworlds*, pages 67–71. IEEE, 2010.

Harry T Reis, Ilona McDougal Wilson, Carla Monestere, Stuart Bernstein, Kelly Clark, Edward Seidl, Michelle Franco, Ezia Gioioso, Lori Freeman, and Kimberly Radoane. What is smiling is beautiful and good. *European Journal of Social Psychology*, 20(3):259–267, 1990.

Michael J Richardson, Kerry L Marsh, Robert W Isenhower, Justin RL Goodman, and Richard C Schmidt. Rocking together: Dynamics of intentional and unintentional interpersonal coordination. *Human movement science*, 26(6):867–891, 2007.

Lazlo Ring, Timothy Bickmore, and Paola Pedrelli. An affectively aware virtual therapist for depression counseling. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI) workshop on Computing and Mental Health*, pages 01951–12, 2016.

Hannes Ritschel, Tobias Baur, and Elisabeth André. Adapting a robot's linguistic style based on socially-aware reinforcement learning. In *2017 26th ieee international symposium on robot and human interactive communication (ro-man)*, pages 378–384. IEEE, 2017.

Stuart J Russell. *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.

Najmeh Sadoughi and Carlos Busso. Novel realizations of speech-driven head movements with generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6169–6173. IEEE, 2018.

Takeshi Saga, Hiroki Tanaka, Yasuhiro Matsuda, Tsubasa Morimoto, Mitsuhiro Uratani, Kosuke Okazaki, Yuichiro Fujimoto, and Satoshi Nakamura. Automatic evaluation-feedback system for automated social skills training. *Scientific Reports*, 13(1):6856, 2023a.

Takeshi Saga, Jieyeon Woo, Alexis Gerard, Hiroki Tanaka, Catherine Achard, Satoshi Nakamura, and Catherine Pelachaud. An adaptive virtual agent platform for automated social skills training. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2023b.

Murat Şahin and Eren Aybek. Jamovi: an easy to use statistical software for the social scientists. *International Journal of Assessment Tools in Education*, 6(4): 670–692, 2019.

Jörn PW Scharlemann, Catherine C Eckel, Alex Kacelnik, and Rick K Wilson. The value of a smile: Game theory with a human face. *Journal of Economic Psychology*, 22(5):617–640, 2001.

Klaus R Scherer, Rainer Banse, Harald G Wallbott, and Thomas Goldbeck. Vocal cues in emotion encoding and decoding. *Motivation and emotion*, 15:123–148, 1991.

Iony D Schmidt, Benjamin J Pfeifer, and Daniel R Strunk. Putting the "cognitive" back in cognitive therapy: Sustained cognitive change as a mediator of in-session insights and depressive symptom improvement. *Journal of Consulting and Clinical Psychology*, 87(5):446, 2019.

Richard C Schmidt and Michael J Richardson. Dynamics of interpersonal coordination. In *Coordination: Neural, behavioral and social dynamics*, pages 281–308. Springer, 2008.

Marc Schroder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark Ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, et al. Building autonomous sensitive artificial listeners. *IEEE transactions on affective computing*, 3(2):165–183, 2011.

Dhruv Sharma, Sanjay Purushotham, and Chandan K Reddy. Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain. *Scientific Reports*, 11(1):19826, 2021.

Christopher F Sharpley, Jennifer Halat, Tammy Rabinowicz, Birgit Weiland, and Jane Stafford. Standard posture, postural mirroring and client-perceived rapport. *Counselling Psychology Quarterly*, 14(4):267–280, 2001.

Youssef Shiban, Iris Schelhorn, Verena Jobst, Alexander Hörnlein, Frank Puppe, Paul Pauli, and Andreas Mühlberger. The appearance effect: Influences of virtual agent features on performance and motivation. *Computers in Human Behavior*, 49:5–11, 2015.

Kazuhiro Shidara, Hiroki Tanaka, Hiroyoshi Adachi, Daisuke Kanayama, Yukako Sakagami, Takashi Kudo, and Satoshi Nakamura. Automatic thoughts and facial expressions in cognitive restructuring with virtual agents. *Frontiers in Computer Science*, 4:8, 2022.

Kevin Shockley, Marie-Vee Santana, and Carol A Fowler. Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance*, 29(2):326, 2003.

Candace L Sidner, Christopher Lee, and Neal Lesh. Engagement when looking: behaviors for robots when collaborating with people. In *Diabruck: Proceedings of the 7th workshop on the Semantic and Pragmatics of Dialogue*, pages 123–130. Citeseer, 2003.

Candace L Sidner, Timothy Bickmore, Bahador Nooraie, Charles Rich, Lazlo Ring, Mahni Shayganfar, and Laura Vardoulakis. Creating new technologies for companionable agents to support isolated older adults. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(3):1–27, 2018.

Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.

Bruce Snyder, Dejan Bosnanac, and Rob Davies. *ActiveMQ in action*, volume 47. Manning Greenwich Conn., 2011.

Charles D Spielberger, Fernando Gonzalez-Reigosa, Angel Martinez-Urrutia, Luiz FS Natalicio, and Diana S Natalicio. The state-trait anxiety inventory. *Revista Interamericana de Psicologia/Interamerican journal of psychology*, 5(3 & 4), 1971.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Yongxue Tian and Li Pan. Predicting short-term traffic flow by long short-term memory recurrent neural network. In *2015 IEEE international conference on smart city/SocialCom/SustainCom (SmartCity)*, pages 153–158. IEEE, 2015.

Linda Tickle-Degnen and Robert Rosenthal. The nature of rapport and its nonverbal correlates. *Psychological inquiry*, 1(4):285–293, 1990.

Larissa Z Tiedens and Alison R Fragale. Power moves: complementarity in dominant and submissive nonverbal behavior. *Journal of personality and social psychology*, 84(3):558, 2003.

Khiet P Truong, Ronald Poppe, and Dirk Heylen. A rule-based backchannel prediction model using pitch and pause information. In *Eleventh Annual Conference of the International Speech Communication Association*. Citeseer, 2010.

Avraam Tsantekidis, Nikolaos Passalis, Anastasios Tefas, Juho Kanniainen, Moncef Gabbouj, and Alexandros Iosifidis. Forecasting stock prices from the limit order book using convolutional neural networks. In *2017 IEEE 19th conference on business informatics (CBI)*, volume 1, pages 7–12. IEEE, 2017.

Wolfgang Tschacher, Georg M Rees, and Fabian Ramseyer. Nonverbal synchrony and affect in dyadic interactions. *Frontiers in psychology*, 5:1323, 2014.

Nguyen Tan Viet Tuyen and Oya Celiktutan. Context-aware human behaviour forecasting in dyadic interactions. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, pages 88–106. PMLR, 2022.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Jelte van Waterschoot, Merijn Bruijnes, Jan Flokstra, Dennis Reidsma, Daniel Davison, Mariët Theune, and Dirk Heylen. Flipper 2.0: A pragmatic dialogue engine for embodied conversational agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 43–50, 2018.

Gary R VandenBos. *APA dictionary of psychology*. American Psychological Association, 2007.

Giovanna Varni, Gualtiero Volpe, and Antonio Camurri. A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media. *IEEE Transactions on Multimedia*, 12(6):576–590, 2010.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Lisa N Vittorio, Samuel T Murphy, Justin D Braun, and Daniel R Strunk. Using socratic questioning to promote cognitive change and achieve depressive symptom reduction: evidence of cognitive change as a mediator. *Behaviour research and therapy*, 150:104035, 2022.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.

Astrid M Von der Pütten, Nicole C Krämer, Jonathan Gratch, and Sin-Hwa Kang. "it doesn't matter what you are!" explaining social effects of agents and avatars. *Computers in Human Behavior*, 2010.

Renzhuo Wan, Shuping Mei, Jun Wang, Min Liu, and Fan Yang. Multivariate temporal convolutional network: A deep neural networks approach for multivariate time series forecasting. *Electronics*, 8(8):876, 2019.

Isaac Wang and Jaime Ruiz. Examining the use of nonverbal communication in virtual agents. *International Journal of Human–Computer Interaction*, 37(17): 1648–1673, 2021.

Ning Wang and Jonathan Gratch. Can virtual human build rapport and promote learning? In *Artificial Intelligence in Education*, pages 737–739. IOS Press, 2009.

Klaus Weber, Hannes Ritschel, Ilhan Aslan, Florian Lingenfelser, and Elisabeth André. How to shape the humor of a robot-social behavior adaptation based on reinforcement learning. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 154–162, 2018.

Jieyeon Woo. Development of an interactive human/agent loop using multimodal recurrent neural networks. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 822–826, 2021.

Jieyeon Woo, Catherine Pelachaud, and Catherine Achard. Creating an interactive human/agent loop using multimodal recurrent neural networks. In *WACAI 2021*, 2021.

Jieyeon Woo, Mireille Fares, Catherine Pelachaud, and Catherine Achard. Amii: Adaptive multimodal inter-personal and intra-personal model for adapted behavior synthesis. *arXiv preprint arXiv:2305.11310*, 2023a.

Jieyeon Woo, Michele Grimaldi, Catherine Pelachaud, and Catherine Achard. Conducting cognitive behavioral therapy with an adaptive virtual agent. In *ACM International Conference on Intelligent Virtual Agents (IVA '23)*, 2023b.

Jieyeon Woo, Michele Grimaldi, Catherine Pelachaud, and Catherine Achard. Iava: Interactive and adaptive virtual agent. In *ACM International Conference on Intelligent Virtual Agents (IVA '23)*, 2023c.

Jieyeon Woo, Catherine Pelachaud, and Catherine Achard. Asap: Endowing adaptation capability to agent in human-agent interaction. In *28th International Conference on Intelligent User Interfaces*, 2023d.

Jieyeon Woo, Catherine Pelachaud, and Catherine Achard. Reciprocal adaptation measures for human-agent interaction evaluation. In *ICAART*, 2023e.

Jieyeon Woo, Kazuhiro Shidara, Catherine Achard, Hiroki Tanaka, Satoshi Nakamura, and Catherine Pelachaud. Adaptive virtual agent: Design and evaluation for real-time human-agent interaction. *International Journal of Human-Computer Studies*, 2023f.

Jieyeon Woo, Liu Yang, Catherine Achard, and Catherine Pelachaud. Are we in sync during turn switch? In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–4. IEEE, 2023g.

Jieyeon Woo, Liu Yang, Catherine Pelachaud, and Catherine Achard. Is turn-shift distinguishable with synchrony? In *International Conference on Human-Computer Interaction*, pages 419–432. Springer, 2023h.

Matthew Wright. Open sound control: an enabling technology for musical networking. *Organised Sound*, 10(3):193–200, 2005.

Haimin Yang, Zhisong Pan, Qing Tao, et al. Robust and adaptive online time series prediction with long short-term memory. *Computational intelligence and neuroscience*, 2017, 2017.

Hani C Yehia, Takaaki Kuratate, and Eric Vatikiotis-Bateson. Linking facial animation, head motion and speech acoustics. *Journal of phonetics*, 30(3):555–568, 2002.

Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, pages 346–364. Springer, 2020.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.