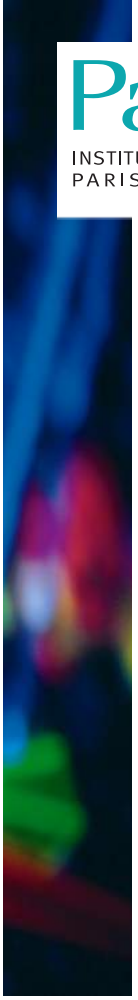
Agents Conversationnels Animés pour l'entrainement social : modèle computationnel de l'expression d'attitudes sociales par des séquences de signaux non-verbaux

Mathieu Chollet



ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2015-ENST-0017

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Traitement du Signal et des Images »

présentée et soutenue publiquement par

Mathieu CHOLLET

le 21 avril 2015

Agents Conversationnels Animés pour l'entrainement social : modèle computationnel de l'expression d'attitudes sociales par des séquences de signaux non-verbaux

Directeur de thèse : **Catherine PELACHAUD**
Co-encadrement de la thèse : **Magalie OCHS**

Jury

M. Pierre DE LOOR, Professeur, LABSTICC, ENIB	Président
M. Dirk HEYLEN, Professeur, HMI - CTIT, Université de Twente	Rapporteur
M. Jean-Claude MARTIN, Professeur, LIMSI-CNRS, Université Paris Sud	Rapporteur
Mme Domitile LOURDEAUX, Maître de Conférence, Heudiasyc, UTC	Examinatrice
M. Thierry PUN, Professeur, CVML, Université de Genève	Examinateur
M. Nicolas SABOURET, Professeur, LIMSI-CNRS, Supélec	Invité
Mme Magalie OCHS, Maître de Conférence, LSIS, Unversité d'Aix-Marseille	Encadrante de thèse
Mme Catherine PELACHAUD, Directeur de recherche, LTCI, Telecom Paristech	Directrice de thèse

TELECOM ParisTech
école de l'Institut Mines-Télécom - membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

T H È S E

Résumé

Les développements récents dans le domaine des Agents Conversationnels Animés (ACAs), des personnages virtuels utilisant les modalités de la communication humaine (*ex.* gestes, parole) pour interagir avec un utilisateur, ont permis de proposer des systèmes pour l'entrainement des compétences sociales. Ces ACAs ont d'abord pris le rôle de tuteurs, ou de compagnons d'apprentissage, mais de nouvelles approches proposent l'utilisation des personnages virtuels afin de pratiquer ces compétences sociales en situation. Le projet Tardis propose par exemple d'utiliser un ACA dans le rôle d'un recruteur virtuel afin de s'entrainer à passer des entretiens d'embauche. Les contributions de cette thèse concernent deux enjeux de la conception de systèmes d'entrainement social par la mise en situation à l'aide d'ACAs.

Les ACAs utilisés dans l'entrainement social doivent être capables de confronter l'apprenant à toutes les situations sociales nécessaires à son apprentissage. Selon l'application, les ACAs doivent ainsi pouvoir exprimer différentes émotions ou différentes attitudes. Les émotions et attitudes s'expriment en particulier par les signaux non-verbaux : sourires, postures, *etc.* Des recherches en psychologie ont cependant montré que l'interprétation de signaux non-verbaux est modifiée par les signaux proches : un sourire suivi d'un évitement de regard et d'un détournement de la tête n'est ainsi pas un signe d'amusement mais d'embarras. Lors de la planification des signaux non-verbaux d'un ACA, il faut donc considérer les séquences complètes de signaux non-verbaux et pas seulement ces signaux indépendamment. Or, la plupart des modèles existants ne prennent pas cela en compte, ou de manière limitée. La contribution principale de notre thèse est de proposer une méthodologie pour extraire automatiquement d'un corpus multimodal des séquences de signaux non-verbaux caractéristiques d'un phénomène socio-émotionnel étudié, et un modèle de planification de signaux non-verbaux utilisant un ensemble de séquences extraites avec cette méthode. Nous appliquons ces travaux à l'expression d'attitudes sociales par un recruteur virtuel.

Une autre enjeu dans la conception de systèmes d'entrainement social est de vérifier si les utilisateurs améliorent bien leurs compétences sociales en s'entrainant avec de tels systèmes. Nous nous sommes intéressés à la réalisation d'une audience virtuelle constituée d'ACAs servant à s'entrainer à l'entrainement à la prise de parole en public. Notre seconde contribution est d'avoir proposé une architecture d'audience virtuelle réagissant en temps réel à la performance de l'utilisateur, et d'avoir évalué plusieurs stratégies de retour à l'utilisateur.

Remerciements

Avec la rédaction de ce manuscrit de thèse, c'est une période de ma vie riche en rencontres, découvertes et bouleversements qui s'achève, et la tentation de faire le bilan me vient tout naturellement, c'est le cas de tout doctorant j'imagine.

On m'avait bien sûr prévenu qu'une thèse était un parcours jalonné de hauts et de bas, mais comme pour de nombreuses choses dans la vie, il faut le vivre pour s'en rendre compte. On y trouve pèle-mêle des réussites ou des échecs, des grands changements, des questionnements sur soi-même, ce qu'on était, ce que l'on est et ce que l'on devient, des voyages, des souvenirs, des déceptions. C'est aussi beaucoup de temps qui passe sans que l'on s'en rende compte, sans qu'on ait l'impression que quoi que ce soit n'avance ou bien ne change, même si c'est le cas, en lame de fond, quelque part en soi. Et d'un coup, c'est la fin, tout se termine en un éclair !

Après avoir pris le temps du repos, après les derniers instants exténuants du parcours du thésard, je prends ici le temps de diriger mes pensées vers les personnes qui, de près ou de loin, m'ont accompagné dans ce chemin accidenté.

Je voudrais tout d'abord remercier les membres de mon jury. Je m'estime excessivement chanceux d'avoir pu être jugé par un groupe de chercheurs constitué de personnes si talentueuses dans leurs recherches, mais aussi si chaleureuses humainement. Je voudrais ainsi remercier Dirk Heylen et Jean-Claude Martin, qui ont acceptés d'officier en tant que rapporteurs de cette thèse, pour l'intérêt et la bienveillance dont ils ont fait preuve. Je voudrais également remercier Domitile Lourdeaux et Thierry Pun, qui ont pris part à ma soutenance de thèse en tant qu'examinateurs, et Pierre De Loor (qui fût, alors que j'étais encore en Master à Telecom Bretagne, un de mes premiers contacts avec le monde de la recherche - la boucle est bouclée) qui l'a présidée. Merci enfin à Nicolas Sabouret, qui aura été à la fois un membre invité de mon jury de thèse, un encadrant (mais pas vraiment), un collègue (mais pas vraiment non plus), et un ami. Grâce à toutes ces personnes, le moment (tant redouté) de la séance de questions a finalement été un plaisir, grâce à la profondeur de leur analyse et à leur lecture attentionnée de mon manuscrit, ainsi qu'à leurs nombreux compliments, dont je ne m'estimerai probablement jamais totalement digne et dont

je me suis senti profondément honoré.

Cette aventure et cette belle conclusion ont d'abord pu avoir lieu grâce à Catherine Pelachaud et Magalie Ochs, mes deux directrices de thèse. Il est difficile de décrire à quel point leur présence et leur accompagnement, perpétuellement bienveillant, à été crucial pour le bon déroulement de ma thèse. Après quelques débuts hésitants, elles ont su me placer sur des rails solides et s'assurer ensuite que je ne m'égare pas. Tout au long de ce processus, j'ai pu me rendre compte que cet engagement que vous avez manifesté à mon égard et envers vos autres doctorants n'est pas seulement guidé par une responsabilité du professeur envers l'élève, ou par un intérêt commun pour la recherche, mais bel et bien par une réelle amitié, et par un vrai attachement à notre bien-être et à notre bonheur personnel. Je ne peux ici trouver des mots pour exprimer à quel point cela est précieux.

J'ai eu la chance de réaliser ma thèse dans un laboratoire qui ne manque pas de bonne humeur, et dont les membres doivent commencer à trouver que le temps est bien long dans ces remerciements sans que l'on parle d'eux (je t'ai vu, Nesrine !). Merci donc à toute l'équipe Greta/Greto/DB302 pour tous ces moments. Des remerciements à toute l'équipe, en particulier à Brian et Florian, pour les voyages, les surnoms et tout le reste, à Nesrine, ma « soeur de thèse », à Sylwia pour son écoute, mais aussi à Angelo, Kevin, Brice, Caroline, Nadine, Sabrina, Beatrice, André-Marie, Thomas, Amyr, Ken, Chloé, les deux Pierre, Jessica, Ling, Nick, Floriane, Abhishek, Yu, Jing, Matthieu, Clémence, Soumia, et Irina.

Merci aussi aux autres personnes que j'ai côtoyées au cours de ma thèse et qui ont, de près ou de loin, soit aidé à la réaliser, soit aidé à rendre le voyage plus agréable : aux autres membres du projet Tardis, en particulier Hazaël, Kaska, Ionut, Arjan, Cathy, Patrick, Aurélie, Atef, Evi, Norbert, grâce à qui les nombreux séjours studieux en Europe ont toujours été un véritable plaisir. Merci à Candy Sidner pour ses nombreux conseils au début de ma thèse.

J'ai eu la chance de traverser l'Atlantique pendant ma thèse pour rejoindre un autre laboratoire sur les côtes californiennes. La-bas, je voudrais d'abord remercier Stefan Scherer, qui non seulement m'a permis de travailler sur un projet qui me passionne, mais qui est aussi devenu un très bon ami. Merci aussi à Louis-Philippe Morency, pour son accueil en Californie et ses conseils pendant mon séjour à l'ICT.

De Californie je suis aussi reparti plus riche de nombreux amis, à qui même s'ils

vivent tous loin de moi j'aimerais adresser des remerciements. Merci donc à Camille, Shannon, David, Aisha, Trevor, Léa, Pippa, Elnaz, Haley, Philip, Eda, Mikkel, Philip et Maike.

Merci aux amis d'ici grâce à qui ces quelques années si bouleversantes et tumultueuses restent avant tout un bon souvenir. Merci aux Poc, hellfesteux, amateurs de nanards et affiliés, Tanguy, Thomas, Ivan, Guillaume, Romain, Vincent, PY, Nikita, Parpaing, Tus, Anne-Cé, Manue, Isa, Madeleine, Eve. Merci aux Narbonnais, toujours là depuis le temps, Brice, David, Hélène, Babeth, Laura. Merci à Ophélie, Franck et Gilles pour les bons moments en tournée, à Saint-Ouen ou à Paris. Merci à Mathieu et Max pour les souvenirs à Rennes et avec Clare. Merci à tous les amis du labo d'à côté, Cristina, Olivier, Marc, les Sylvain, les Émilie, Paul, Beppe, Eric, Guillaume.

Enfin, je voudrais remercier ceux pour qui tous les mots, tous les espaces, toutes les paroles et tous les silences, ne suffisent pas à exprimer ce que je ressens à leur égard. Merci à Mathilde pour ce qui fera partie de moi à jamais. Merci à ma famille, en particulier mes parents, mon frère, d'être toujours là et de croire en moi dur comme fer. Merci à Gaëlann pour tout ce que nous avons vécu et tout ce qu'il reste à écrire.

Table des matières

Table des figures

Liste des tableaux

1 Introduction

Dans la communication humaine, nous produisons des gestes, des expressions faciales, adoptons certaines postures, parlons avec une prosodie particulière [Knapp et al., 2013] : chacune de ces modalités envoie des signaux, que ce soit consciemment ou non, qui peuvent être interprétés pour déduire nos intentions et nos états mentaux. Les Agents Conversationnels Animés (ACAs) sont un type d'interface multimodale reproduisant ces modalités naturelles de la communication humaine, comme la parole, les gestes, le regard, les postures, les expressions faciales, l'intonation, afin de reproduire l'expérience d'une interaction entre deux humains [Cassell et al., 2000]. On peut remarquer que des ACAs ont été intégrés dans des sites web commerciaux afin de gérer des requêtes clients, à l'instar des agents Léa de la SNCF[1] ou Anna d'Ikea[2]. Des ACAs ont aussi été déployés dans des salles d'expositions à taille réelle comme l'agent Gloria de Charamel[3], ou dans des musées en tant qu'agents d'accueil ou d'installations à part entière, comme les sœurs jumelles Ada et Grace du Boston Museum of Science [Swartout et al., 2010] (voir Figure 1.1).

1.1 L'entrainement virtuel avec des Agents Conversationnels Animés

Les Agents Conversationnels Animés ont aussi été utilisés dans le domaine des environnements informatiques pour l'apprentissage humain. Un Agent Conversationnel Animé peut y endosser le rôle d'un expert, d'un guide [Johnson et al., 2000, Graesser et al., 2003] ou encore d'un pair [Kim & Baylor, 2006] qui va apporter un support à l'utilisateur dans son apprentissage. Un exemple précurseur dans ce domaine est l'agent Steve [Johnson et al., 2000], qui était utilisé dans le cadre de l'apprentissage

[1] http://aide.voyages-sncf.com/
[2] http://www.ikea.com/fr/fr/
[3] http://www.charamel.com/

FIGURE 1.1: Exemples d'Agents Conversationnels Animés. De gauche à droite et de haut en bas : l'agent Léa de la SNCF[1], Gloria, agent de Charamel[3], l'agent Anna d'IKEA[2], et Ada et Grace, agents du Boston Museum of Science [Swartout et al., 2010].

de tâches procédurales pour la maintenance et l'opération de machines (voir Figure 1.2). Steve était représenté dans un environnement virtuel comprenant des modèles 3D des machines dont l'utilisateur devait apprendre le fonctionnement. Il pouvait expliquer à l'apprenant les différentes étapes de certaines tâches en les réalisant dans l'environnement virtuel et en utilisant la parole pour expliquer ses actions ou répondre aux questions. De plus, il pouvait utiliser son comportement non-verbal (*ex.* regard, gestes) pour guider l'attention de l'utilisateur vers des objets importants.

Plus récemment, l'utilisation des Agents Conversationnels Animés pour l'entrainement de compétences sociales par la pratique a été mise en avant. Dans ces situations, un autre rôle émerge pour les Agents Conversationnels Animés : ils peuvent constituer l'objet de pratique en lui-même [Lane & Wray, 2012]. Le rôle de l'ACA n'est alors plus d'apporter un support à un apprentissage traditionnel en tant que guide, expert ou pair, mais de permettre une mise en pratique qui va être source d'apprentissage [Gratch & Marsella, 2005]. Des applications sont apparues dans des

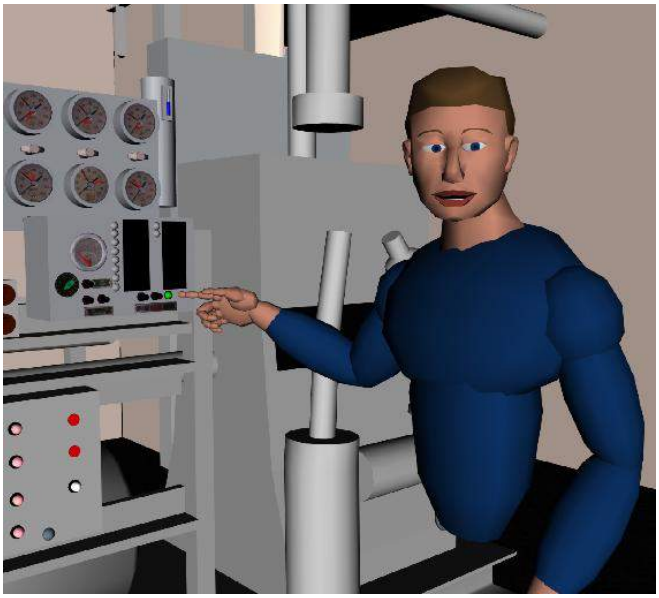


FIGURE 1.2: L'Agent Conversationnel Animé pour l'entrainement virtuel Steve, en train de pointer un objet important par un geste déictique.

domaines variés [Hoque & Picard, 2014], allant de l'entrainement de personnels de santé à délivrer des informations médicales aux patients [Talbot et al., 2012], à l'entrainement des enfants autistes à développer leurs compétences sociales [Bernardini et al., 2012]. Par exemple, dans le cas de l'entrainement de personnels de santé, l'agent va pouvoir prendre le rôle d'un patient virtuel.

Les pratiques traditionnelles pour l'entrainement de compétences sociales, comme l'entrainement à donner une présentation orale ou à passer un entretien d'embauche, consistent à s'entrainer avec des experts ou des proches [Spence, 2003, Lane et al., 2013]. Ces approches posent plusieurs problèmes :

- Une personne désireuse d'améliorer ses compétences sociales peut ne pas vouloir se confronter à l'avis d'autres personnes de peur d'être jugée, ou de se sentir ridicule, lors d'une présentation orale par exemple. En revanche, l'interaction avec des personnages virtuels permet de réduire ces appréhensions [Pertaub et al., 2002, Grillon et al., 2006].

- Le recours à des experts pose des problèmes de disponibilité lorsque ceux-ci sont en nombre insuffisant pour couvrir la demande : c'est le cas dans des associations locales pour l'insertion où de nombreuses personnes en recherche d'emploi sont désireuses de s'entrainer à passer des entretiens d'embauche,

mais où le nombre de conseillers impose de limiter le temps que ceux-ci peuvent accorder à chaque personne. De la même manière, des amis ou des collègues ne sont pas toujours disponibles pour participer à un entrainement. Inversement, la seule limite à la disponibilité d'un logiciel est celle de l'accès initial à cette technologie.

- Il n'est pas possible de standardiser complètement les interventions ou mises en situation pour l'entrainement de compétences sociales, de par la variabilité des situations sociales, des relations entre l'apprenant et ses partenaires d'apprentissages, voire de leurs humeurs. Dans le cas d'un logiciel, toutes les variables d'une interaction peuvent être précisément contrôlées.

Des applications d'entrainement des compétences sociales utilisant des Agents Conversationnels Animés peuvent apporter une solution à ces problèmes. Cependant, la conception de tels systèmes comporte plusieurs enjeux majeurs.

1.2 Enjeux dans la conception d'Agents Conversationnels Animés pour l'entrainement social

Dans notre travail de thèse, nous nous sommes concentrés sur deux enjeux de la conception d'ACAs pour l'entrainement social. Le premier concerne la planification du comportement d'Agents Conversationnels Animés pour l'expression d'attitudes. Le deuxième concerne la conception et l'évaluation de systèmes dédiés à l'amélioration de la compétence de prise de parole en public.

1.2.1 Planification du comportement non-verbal d'Agents Conversationnels Animés pour l'expression d'attitudes

Selon l'application envisagée, la conception des Agents Conversationnels Animés peut être plus ou moins complexe. Par exemple, le comportement d'un ACA peut être prévu à l'avance (scripté) lors d'interactions totalement contrôlées (*ex.* interaction avec l'ACA par le biais d'un menu dans un scénario fixe). Inversement, si l'objectif est de simuler une interaction sociale complète entre un utilisateur et un ACA autonome, alors la tâche est beaucoup plus complexe. Il faut doter cet ACA de capacités de reconnaissance des signaux, gestes, phrases de l'utilisateur ; il doit

pouvoir participer à la régulation de la discussion : par exemple, en trouvant le bon moment où prendre la parole ; enfin, selon l'application, il doit être capable d'exprimer des états socio-émotionnels (*ex.* de l'empathie, des émotions, des attitudes) de façon opportune et en s'assurant que leur expression est bien reconnue par l'utilisateur. Dans un entretien d'embauche, un candidat va être confronté à des recruteurs plus ou moins amicaux, plus ou moins dominants, c'est à dire des recruteurs exprimant différentes attitudes sociales. Quelque soit l'attitude du recruteur, le candidat doit pouvoir faire la meilleure impression possible. Dans le cadre de systèmes d'entrainement aux entretiens d'embauche, il est donc essentiel de pouvoir exprimer les attitudes de recruteurs virtuels, afin de confronter des utilisateurs aux différentes attitudes que pourraient adopter de réels recruteurs.

Ces fonctions communicatives que nous venons d'évoquer sont assurées par les signaux multimodaux (*ex.* gestes, expressions faciales, regard, *etc.*). Les modèles de planification de comportement d'ACAs doivent donc prendre en compte de nombreuses fonctions et considérer la manière dont sont interprétés les signaux multimodaux qui peuvent les remplir. L'approche standard dans le domaine de la planification de comportement d'ACAs consiste à définir un lexique dans lesquels on spécifie les signaux qui peuvent être utilisés pour réaliser les différentes fonctions communicatives considérées [Cassell et al., 2004, Poggi et al., 2005, Lee & Marsella, 2006, Hartholt et al., 2013, Cafaro, 2014]. Lorsque des fonctions communicatives doivent être réalisées par l'ACA, il suffit alors de se référer à ce lexique pour sélectionner les signaux non-verbaux appropriés pour chacune de ces fonctions. Pour définir un tel lexique, on peut utiliser la littérature en sciences humaines et sociales [Ekman & Friesen, 1976, Kendon, 1967, Poggi, 2003, Knapp et al., 2013], analyser des corpus multimodaux [Martin et al., 2006, Niewiadomski et al., 2011] ou ecore développer des modèles d'apprentissage [Deng & Neumann, 2008, Ding et al., 2014]. Par exemple, les expressions faciales d'émotions ont longtemps été étudiées par le biais de photographies montrant une personne au pic de l'expression d'une émotion [Ekman & Friesen, 1976, Ekman, 2007, Hess et al., 2007]. Pour permettre à un ACA d'exprimer ces émotions, il suffit de reproduire dans un lexique les descriptions des expressions faciales de ces photographies [Pelachaud et al., 1996, Pandzic & Forchheimer, 2002].

Cette approche permet de réaliser chaque fonction communicative spécifiée dans le lexique en des signaux multimodaux. Cependant, ces réalisations sont effectuées

indépendamment les unes des autres, c'est à dire que le choix des signaux utilisés pour réaliser une intention communicative est fait sans prendre en compte les autres intentions communicatives et les autres signaux multimodaux proches. Une telle approche est insuffisante pour modéliser l'expression d'attitudes sociales. Scherer présente une définition des attitudes où il les décrit comme « caractéristiques d'un style affectif qui colore une interaction » (voir Section 2.2.1) [Scherer, 2005]. Une attitude sociale n'est donc pas un phénomène qui s'exprime à un instant précis par une expression prototypique, comme c'est le cas de certaines émotions. Les attitudes se manifestent dans l'ensemble du comportement d'une personne ; elles sont aussi communiquées sur une plus longue période. De plus, des études ont montré que l'interprétation de certains signaux non-verbaux peut être modifiée par d'autres signaux proches [Keltner, 1995, With & Kaiser, 2011, Jack et al., 2014]. Par exemple, un sourire n'est pas forcément un signe d'amusement ou d'amicalité mais peut indiquer de l'embarras s'il est suivi d'un évitement de regard et d'un détournement de la tête [Keltner, 1995]. Une approche de planification de comportement pour l'expression d'attitudes devra donc assurer la cohérence des différents signaux générés pour des intentions communicatives différentes par rapport à l'attitude voulue. A ce jour, aucun modèle d'expression d'attitude ne prend en compte cette cohérence entre des signaux proches.

1.2.2 Conception et évaluation de systèmes d'entrainement de la compétence de prise de parole en public

La capacité à prendre la parole en public est une compétence essentielle dans de nombreuses situations personnelles et professionnelles. Les approches traditionnelles pour améliorer celle-ci posent des problèmes de disponibilité, de standardisation et de réticence à l'entrainement de peur d'être jugé. Un système d'entrainement dédié à l'amélioration de la compétence de prise de parole en public pourrait constituer une solution à ces problèmes. A ce jour, des audiences virtuelles ont été utilisées pour la réduction de l'anxiété de personnes anxieuses lors de prises de parole en public. Cependant, celles-ci n'ont pas été utilisées et évaluées pour l'amélioration de la compétence de prise de parole en public.

La compétence de prise de parole en public est complexe à évaluer, et un orateur doit présenter plusieurs qualités pour être considéré talentueux [Batrinca et al., 2013].

Il doit d'abord savoir structurer une présentation et choisir le contenu de celle-ci. Le comportement non-verbal est aussi crucial : il doit par exemple avoir une voix sans hésitations, qui ne doit pas être monotone ; il doit regarder le public, adopter une bonne posture, occuper un espace adéquat dans la pièce. Un système dédié à l'entrainement de la compétence de prise de parole en public devra pouvoir fournir un retour à l'utilisateur sur ces différents aspects qui, ensemble, constituent sa performance globale. La manière de délivrer ces retours est importante et peut avoir une influence sur l'entrainement. Ces retours peuvent être fournis sous différentes formes, par exemple sous forme de bilan après un entrainement, ou bien par le biais de mesures objectives fournies au cours d'un entrainement. L'utilisation d'audiences virtuelles constituées d'Agents Conversationnels Animés permet de délivrer ces retours par le biais de signaux non-verbaux. Il est important d'évaluer l'amélioration des compétences des participants en fonction de ces types de retours afin de rendre les systèmes d'entrainement des compétences sociales plus efficaces.

Les contributions de notre thèse s'inscrivent dans le cadre de ces deux enjeux. Nous nous sommes attachés à proposer un modèle de planification de comportement permettant à un recruteur virtuel d'exprimer des attitudes. Ce modèle considère des séquences de signaux non-verbaux afin de s'assurer de la cohérence de ceux-ci par rapport à l'attitude exprimée. De plus, nous avons étudié l'amélioration de la compétence de prise de parole en public par l'utilisation d'une audience virtuelle interactive.

1.3 MÉTHODOLOGIE ET CONTRIBUTIONS

Notre première problématique de recherche a concerné la planification du comportement d'Agents Conversationnels Animés. Comme nous l'avons présenté dans la section précédente, il est nécessaire de considérer des séquences de signaux non-verbaux et pas seulement des signaux indépendants lors de la planification du comportement d'ACAs pour l'expression d'attitudes. Notre objectif a donc été de proposer un modèle de planification de séquences de signaux non-verbaux. Nous avons appliqué ce modèle à l'expression d'attitudes sociales par un recruteur virtuel. Notre méthodologie et nos contributions ont été les suivantes.

1.3.1 Un recruteur virtuel exprimant des attitudes sociales par des séquences de signaux non-verbaux

De nombreuses études ont été publiées sur le rôle de certaines modalités dans l'expression des attitudes, vues indépendamment des autres modalités [Burgoon et al., 1984, Argyle, 1988, Gifford, 1994, Gifford & Hine, 1994, Burgoon & Le Poire, 1999, Carney et al., 2005, Yabar & Hess, 2007]. Cependant ces travaux ne nous fournissent pas d'informations sur l'influence de séquences de signaux non-verbaux sur l'expression d'attitudes. Nous avons donc décidé d'extraire des connaissances automatiquement à partir d'un corpus multimodal. Ce corpus constitue notre première contribution.

- *Première contribution :* Corpus multimodal d'entretiens d'embauche annotés au niveau des signaux non-verbaux et de l'expression d'attitudes.

Nous avons annoté ce corpus à deux niveaux : les signaux non-verbaux d'une part, et l'attitude sociale d'autre part. Nous avons proposé d'annoter les attitudes de manière continue (voir Section 4.3). En effet, une représentation fréquemment utilisée des attitudes est celle d'Argyle qui les représente sur deux dimensions continues (amicalité et dominance)[Argyle, 1988]. De plus, comme l'indique Scherer, les attitudes ne sont pas exprimées à un instant donné mais colorent tout un échange interpersonnel [Scherer, 2005]. Une représentation continue semble donc la plus naturelle pour représenter les attitudes.

Une fois ce corpus obtenu, nous nous sommes attachés à déterminer une méthodologie permettant d'identifier automatiquement des séquences de signaux non-verbaux caractéristiques de différentes attitudes. La difficulté était ainsi de faire le lien entre les annotations des signaux non-verbaux, codées sous forme d'évènements possédant un début et une fin, et les attitudes, annotées de manière continue. Notre deuxième contribution est d'avoir proposé une méthodologie d'extraction automatique de séquences de signaux non-verbaux.

- *Deuxième contribution :* Méthodologie d'extraction de séquences de signaux non-verbaux caractéristiques de variations d'attitudes.

Cette méthodologie tire profit de la nature continue des annotations d'attitudes afin de détecter automatiquement les instants où les attitudes varient. Une technique de fouille de motifs séquentiels est ensuite utilisée afin d'identifier les séquences

fréquemment observées avant certains types de variations d'attitudes. Nous avons appliqué cette méthodologie à notre corpus multimodal d'entretiens d'embauche, et avons ainsi obtenu une base de données de séquences de signaux non-verbaux caractéristiques de variations d'attitudes.

L'étape suivante de notre travail a consisté à identifier une méthode permettant d'utiliser cette base de données de séquences de signaux non-verbaux afin de planifier le comportement d'un ACA. En effet, le comportement d'un ACA ne dépend pas que d'une attitude à exprimer, mais aussi de fonctions communicatives à réaliser (*ex.* poser une question, indiquer que l'on désire prendre la parole : voir Section 2.1). Les séquences extraites précédemment n'étaient ainsi pas utilisables directement. Nous avons alors proposé un modèle de planification de séquences de signaux non-verbaux pour l'expression d'attitudes qui tire profit de cette base de données de séquences de signaux non-verbaux. Ce modèle constitue notre troisième contribution.

- *Troisième contribution :* Modèle de planification de séquences de signaux non-verbaux pour l'expression d'attitudes.

Ce modèle de planification de séquences de signaux non-verbaux s'intègre dans l'architecture standard d'Agents Conversationnels Animés SAIBA [Vilhjálmsson et al., 2007, Heylen et al., 2008]. Celui-ci permet d'exprimer des attitudes tout en exprimant les autres fonctions communicatives du discours.

Nous avons implémenté un recruteur virtuel autonome à partir de notre modèle de planification de signaux non-verbaux et d'autres composants existants. Ce recruteur virtuel exprime des attitudes par son discours et son comportement non-verbal, et assure un comportement d'écoute et de régulation du tour de parole. Nous l'avons ensuite intégré à une plate-forme de simulation d'entretiens d'embauche.

- *Quatrième contribution :* Implémentation et évaluation d'un recruteur virtuel autonome.

Nous avons réalisé une étude pour évaluer l'expression d'attitudes par le recruteur virtuel. Dans cette étude, nous avons comparé la perception des attitudes exprimées par le comportement verbal uniquement, le comportement non-verbal uniquement, ou la combinaison des deux. De plus, nous avons étudié si une différence existe dans les taux de reconnaissance des attitudes entre des participants observant des vidéos du recruteur virtuel, et des participants interagissant avec le recruteur dans le cadre

d'une simulation d'entretien d'embauche.

Notre deuxième problématique de recherche concernait la conception et l'évaluation de systèmes d'amélioration de la compétence de prise de parole en public.

1.3.2 Conception et évaluation d'une audience virtuelle interactive

Des audiences virtuelles ont été proposées pour aider des personnes souffrant d'anxiété lors de situations de prise de parole en public à réduire cette anxiété par la pratique. Les audiences virtuelles pourraient aussi constituer un outil pour améliorer la compétence de prise de parole en public : cependant, celles-ci n'ont pas encore été étudiées dans ce contexte. Nous avons proposé et implémenté une architecture d'audience virtuelle interactive pour l'amélioration de la compétence de prise de parole en public.

- *Cinquième contribution :* Architecture et implémentation d'une audience virtuelle interactive.

Le principe de ce système est d'analyser automatiquement certains aspects du comportement constituant une bonne prise de parole en public (*ex.* regarder le public, parler sans hésitations), et de faire un retour en temps réel à l'utilisateur sur ces aspects de sa performance par le biais du comportement des personnages constituant l'audience virtuelle. Nous avons ensuite évalué l'impact de différentes stratégies de retour à l'utilisateur sur l'amélioration de ses compétences.

- *Sixième contribution :* Comparaison de stratégies de retours en temps réel à l'utilisateur.

Nous avons comparé une audience virtuelle passive (*i.e.* affichant uniquement une animation neutre) à une audience virtuelle passive enrichie d'éléments de retour direct (*i.e.* barre de performance), à une audience virtuelle interactive, fournissant un retour à l'utilisateur par son comportement non-verbal (*ex.* se pencher en avant quand la performance de l'utilisateur est bonne, se pencher en arrière en croisant les bras quand elle est mauvaise). Nous avons évalué ces configurations par le biais de trois types de mesures : des questionnaires d'auto-évaluation remplis par les participants, des évaluation par des experts, et des mesures objectives.

1.4 CONTEXTE DE LA THÈSE

Cette thèse a été réalisée au laboratoire CNRS-LTCI de Télécom Paristech, école de l'institut Mines-Télécom, et s'est déroulée dans le cadre du projet européen FP7 Tardis (*Training young Adult's Regulation of emotions and Development of social Interaction Skills*), dont le but est de produire une plate-forme d'entrainement des compétences sociales, fondée sur une simulation d'entretien d'embauche, à destination de jeunes adultes en situation de recherche d'emploi. L'idée novatrice principale du projet Tardis est de fournir une boucle d'interaction en temps réel complète, allant de la reconnaissance du comportement non-verbal de l'utilisateur, à son interprétation en termes de performance, au calcul d'une attitude à exprimer par le recruteur virtuel en réponse à cette performance, et à l'expression de cette attitude par le recruteur virtuel (voir Figure 7.1). Dans le contexte de ce projet, nous avons développé et intégré l'ACA prenant le rôle du recruteur dans les simulations d'entretiens d'embauche.

Au cours de notre thèse, nous avons aussi réalisé une collaboration avec l'Institute for Creative Technologies (ICT) de l'Université de Californie du Sud. Le but de cette collaboration était d'étendre un prototype d'audience virtuelle dédié à l'entrainement des compétences sociales afin de rendre l'audience virtuelle interactive, c'est à dire réagissant automatiquement à la performance de l'utilisateur. Dans ce contexte, nous avons proposé et implémenté une architecture d'audience virtuelle interactive. Nous avons étudié plusieurs stratégies de retour à l'utilisateur sur sa performance afin de déterminer lesquelles sont les plus efficaces en termes d'amélioration de la compétence des utilisateurs.

1.5 ORGANISATION DE CE DOCUMENT

Cette thèse est organisée de la manière suivante. Tout d'abord, nous présentons dans le *chapitre 2* des fondements théoriques sur les signaux non-verbaux et leurs fonctions. Nous y détaillons en particulier le concept d'attitude sociale et les signaux non-verbaux qui participent à son expression. Le *chapitre 3* présente un état de l'art des recherches dans le domaine des Agents Conversationnels Animés, orienté selon trois thématiques centrales de notre thèse : les ACAs et l'entrainement social, les ACAs et l'expression de l'attitude, et la planification de séquences de signaux non-verbaux

pour des ACAs. Le *chapitre 4* est dédié au corpus multimodal qui a été utilisé pour étudier l'expression d'attitudes par des séquences de signaux non-verbaux. Ensuite, nous introduisons dans le *chapitre 5* une méthodologie pour l'extraction automatique de séquences de signaux non-verbaux caractéristiques de certaines attitudes. A partir des séquences extraites de notre corpus multimodal avec cette méthode, nous construisons un modèle de planification de séquences de signaux non-verbaux. Celui-ci est détaillé dans le *chapitre 6*. Le *chapitre 7* présente l'architecture et l'implémentation du recruteur virtuel autonome que nous avons proposé et évalué. Le *chapitre 8* est quant à lui dédié au système d'audience virtuelle interactive pour l'entrainement de la compétence de prise de parole en public que nous avons proposé et évalué. Enfin, dans le *chapitre 9*, nous concluons notre thèse en présentant les limites de nos travaux, et en dressant des perspectives de travaux futurs.

2 Fondements théoriques

Une équation issue des travaux de Mehrabian a souvent été utilisée pour avancer que les quantités d'information fournies par le comportement verbal, vocal et non-verbal, sont respectivement de 7%, 38%, 55% [Mehrabian, 1981]. Affirmer que ces proportions sont les mêmes dans toute situation est une généralisation abusive : l'étude ayant mené à ces résultats avait en effet été réalisée dans un contexte très spécifique (dans cette étude, des participants devaient indiquer le degré d'amicalité qu'ils percevaient dans un enregistrement d'une voix féminine prononçant quelques mots, accompagné d'une photo.). Nous pouvons toutefois retenir un résultat essentiel de ces travaux : la communication humaine ne se limite pas au langage parlé. Lors d'interactions, nous produisons des gestes, des expressions faciales, adoptons différentes postures, regardons notre interlocuteur ou détournons le regard, *etc.* Tous ces signaux non-verbaux sont produits et interprétés de manière intuitive et naturelle, et contribuent à communiquer des informations de diverses natures. Un des arguments en faveur de l'utilisation d'Agents Conversationnels Animés, qui permettent d'émuler les même modalités de communication que les humains, est ainsi qu'ils pourraient permettre une communication naturelle et intuitive avec des utilisateurs.

Un signal non-verbal (*ex.* un haussement de sourcils) peut remplir plusieurs fonctions dans la communication. Inversement une fonction communicative (*ex.* mettre l'accent sur un mot) peut être réalisée par plusieurs signaux non-verbaux. La conception d'un modèle de planification de comportement d'Agents Conversationnels Animés nécessite ainsi de connaître le lien entre les différents signaux non-verbaux et les différentes fonctions du comportement non-verbal. Dans notre travail de thèse, nous nous intéressons à un modèle de planification de comportement pour l'expression d'attitudes dans le cadre de simulations d'entretiens d'embauche avec un recruteur virtuel. Pour concevoir un tel ACA, nous devons ainsi savoir comment certaines fonctions essentielles à la communication, comme la gestion du tour de parole, sont

réalisées par le comportement non-verbal. Nous devons aussi identifier comment les différents signaux non-verbaux participent à l'expression d'attitudes.

Dans ce chapitre, nous présentons les appuis théoriques de notre thèse. Tout d'abord, nous introduisons une sélection de taxonomies des fonctions du comportement non-verbal. En effet, afin d'émuler la communication humaine, le comportement des Agents Conversationnels Animés doit pouvoir assurer ces fonctions. Ensuite, nous définissons le concept d'attitude sociale, et en présentons plusieurs définitions puis la représentation que nous adoptons. Nous détaillons aussi comment les différentes modalités du comportement participent à l'expression d'attitudes. Enfin, nous présentons des travaux qui montrent que l'interprétation du comportement non-verbal peut se faire sous plusieurs perspectives. L'interprétation d'un signal non-verbal peut être influencée par les signaux multimodaux simultanés, précédents ou suivants, par les signaux de l'interlocuteur, et par les tendances comportementales de la personne étudiée.

2.1 Fonctions du comportement non-verbal dans la communication multimodale

Les Agents Conversationnels Animés permettent d'émuler la communication humaine en utilisant les mêmes modalités : la parole, mais aussi les gestes, expressions faciales, postures, *etc.* Afin de concevoir des modèles pour planifier le comportement d'un Agent Conversationnel Animé, il faut ainsi connaitre les différentes fonctions du comportement non-verbal afin de pouvoir déclencher les signaux appropriés au bon moment. Comme l'indique Poggi, on peut s'intéresser à la communication en étudiant soit les types de signaux utilisés (*ex.* les différents gestes, différentes expressions faciales), soit les types d'informations communiquées par le biais de ces signaux [Poggi, 2003]. Nous présentons ici une sélection de travaux ayant abouti à des taxonomies de fonctions communicatives et de signaux non-verbaux.

2.1.1 Ekman et Friesen : le répertoire du comportement non-verbal

Ekman et Friesen ont proposé une taxonomie comprenant cinq grandes catégories de signaux non-verbaux [Ekman & Friesen, 1969]. Ces catégories sont différenciées par

leurs types d'utilisation (*ex.* les signaux de cette catégorie sont ils produits intentionnellement ou non ? Coïncident-ils avec le comportement verbal ?), leurs origines (*i.e.* comment un signal non-verbal s'est-il intégré dans le répertoire communicationnel de l'homme ?), et leurs codages (*i.e.* la forme d'un signal non-verbal est-elle directement liée au sens qu'il communique, ou le lien est-il arbitraire ?).

- La première catégorie est celle des *emblèmes*. Ce sont des signaux non-verbaux qui possèdent une définition directe, conventionnellement admise dans une groupe de personnes donné, et qui peuvent se substituer au langage parlé. Par exemple, dans la culture occidentale, tendre le poing avec le pouce levé est utilisé pour signaler son approbation. Cette catégorie comprend surtout des gestes et expressions faciales, mais n'y est pas exclusivement limitée.

- Les *illustrateurs* sont des signaux non-verbaux accompagnant le discours et qui illustrent ce qui est dit. Des gestes de *bâtons* (*i.e.* gestes simples, rythmiques) peuvent par exemple rythmer le discours ou mettre l'accent sur un mot, tandis qu'un geste *déictique* permet de donner une information spatiale (*ex.* pointer l'objet ou la personne dont on parle). Aux *bâtons* et gestes *déictiques*, Ekman et Friesen ajoutent dans cette catégorie les gestes *idéographiques* (*i.e.* illustration d'une pensée), les *mouvements spatiaux* (*i.e.* illustration d'une relation spatiale), les gestes *kinétographiques* (*i.e.* illustration d'un mouvement corporel) et les gestes *pictographiques* (*i.e.* représentation imagée d'un référent).

- Les *manifestations affectives* (*affect displays*) sont des signaux non-verbaux déclenchés par des réactions émotionnelles. Cette catégorie comprend principalement des expressions faciales, même si Ekman et Friesen indiquent que certains signaux corporels peuvent aussi être considérés comme des manifestations extérieures d'une émotion (*ex.* tremblement lié à la peur).

- Les signaux *régulateurs* permettent la gestion et le maintien de la conversation et de l'interaction avec d'autres personnes. Par exemple, des hochements de tête peuvent être utilisés pour indiquer à son interlocuteur qu'on continue à l'écouter, et qu'il peut continuer à parler.

- Les signaux non-verbaux d'*adaptation* (*adaptors*) servent à satisfaire des besoins corporels (*ex.* se gratter, se recoiffer). Ces signaux sont rarement communicatifs, mais peuvent cependant fournir des informations (*ex.* indiquer de la nervosité).

2.1.2 Cosnier et Vaysse : signaux communicatifs

Cosnier et Vaysse proposent une autre classification des signaux non-verbaux [Cosnier & Vaysse, 1997]. Ils font tout d'abord la différence entre les signaux *extra-communicatifs* et les signaux *communicatifs*. Les signaux *extra-communicatifs* (*Adaptors*) sont des gestes de confort : grattages, manipulations d'objets. Ils ne transmettent pas (à priori : voir Section 2.2.3.3) d'information. Les signaux *communicatifs* sont eux directement impliqués dans l'échange d'informations avec son interlocuteur, et sont à regrouper dans trois sous-catégories :

- Les signaux *synchronisateurs*, qui servent à réguler l'interaction et à « co-piloter l'interaction », c'est à dire à s'assurer que son interlocuteur reçoit et comprend nos énoncés, et à réaliser l'alternance du tour de parole.
- Les signaux *quasi-linguistiques* sont des gestes dont le sens est conventionnellement établi dans une certaine culture et peut être substitué à la parole (les *Emblèmes* d'Ekman et Friesen). Par exemple, placer l'index sur la tempe et opérer des rotations alternées de celui-ci permet de suggérer la folie [Cosnier & Vaysse, 1997].
- Les signaux *co-verbaux* accompagnent le discours et regroupent plusieurs sous-catégories :
  - Les signaux *référentiels* explicitent le discours, soit en le complétant par une information spatiale, comme en pointant du doigt ou de la tête (geste *déictique*), soit en illustrant une caractéristique d'un objet référent du discours par des gestes mimant cette caractéristique, par exemple en indiquant qu'une personne est d'une certaine taille en plaçant la paume de sa main horizontalement à la hauteur correspondante (gestes *illustratifs* ou *iconiques*)
  - Les signaux *expressifs* co-verbaux, regroupant notamment de nombreuses expressions faciales, transmettent des informations affectives. Selon la culture et la situation, les affects ressentis peuvent être inhibés ou non, on peut par exemple cacher de la tristesse par un sourire.
  - Les signaux *paraverbaux* regroupent des mouvements servant à rythmer le discours et à marquer, connecter, structurer ses différentes parties. Par

exemple, les gestes de « bâton », peuvent servir à battre de manière répétée le discours, tandis que les haussements de sourcils peuvent accentuer une partie de celui-ci.

2.1.3 Poggi : le monde, l'identité, l'esprit

Poggi identifie trois classes d'informations qui peuvent être communiquées par un locuteur par le biais du comportement verbal et des différentes modalités du comportement non-verbal [Poggi, 2003] :

- *Informations sur le monde :* Nous pouvons communiquer des informations à propos d'évènements, de personnes ou d'objets (concrets ou abstraits), et des relations spatio-temporelles entre ceux-ci.
- *Informations sur son identité :* De nombreux éléments de notre physique et notre comportement fournissent des informations sur notre âge, notre sexe, notre culture, notre personnalité, *etc.*
- *Informations sur l'esprit :* Nous pouvons aussi communiquer nos sentiments à propos d'autres personnes, objets ou évènements, ou encore les raisons qui nous amènent à en parler.

Au sujet de la dernière catégorie, Poggi a aussi défini une taxonomie des signaux servant à fournir des informations sur son état mental, les *Mind Markers* [Poggi, 2003]. Ces marqueurs, ou signaux, peuvent servir à donner des informations sur ses *objectifs*, sur ses *croyances*, et sur ses *émotions*.

- *Croyances :* Nous pouvons communiquer des informations sur notre degré de certitude (*ex.* hausser les épaules pour indiquer que l'on ne sait pas ou que l'on est incertain), et des informations métacognitives (*ex.* claquer des doigts pour indiquer que l'on est en train d'essayer de se rappeler de quelque chose).
- *Objectifs :* Les marqueurs *performatifs* indiquent l'objectif spécifique d'un acte communicatif (*ex.* on *implore* en penchant la tête, en haussant la partie intérieure des sourcils, et en disant « Je vous implore »). Les marqueurs *métalinguistiques* servent à préciser la structure syntaxique (*ex.* une intonation descendante indique qu'une phrase est terminée) et informationnelle (*ex.* un hochement de tête permet de mettre l'accent sur une partie importante d'une

phrase) du discours, tandis que les marqueurs *méta-discursifs* indiquent les relations entre différentes propositions d'une phrase (*ex.* pencher la tête d'un côté puis de l'autre pour indiquer que deux propositions sont liées, comme avec la locution « d'une part ... d'autre part »). Enfin les marqueurs *méta-conversationnels* permettent de réguler la conversation (*ex.* placer la paume en avant pour indiquer que l'on a pas terminé de parler, regarder son interlocuteur à la fin d'une phrase pour lui donner la parole).

- *Émotions :* La dernière catégorie de marqueurs proposée par Poggi est celle des marqueurs servant à communiquer des émotions. Ces marqueurs concernent toutes les modalités : nous exultons en levant les poings, nous tremblons de peur, et notre visage arbore une grande variété d'expressions faciales.

2.1.4 Argyle : fonctions de la communication corporelle

Dans son ouvrage *Bodily Communication*, Argyle avance que le comportement non-verbal possède cinq fonctions [Argyle, 1988] :

- *Expression d'émotions :* Argyle identifie trois raisons pour lesquelles nous exprimons des émotions. Tout d'abord, certaines émotions déclenchent des réactions physiologiques dont les manifestations non-verbales n'ont pas directement pour but la communication avec autrui (*ex.* l'émotion de dégoût nous fait détourner la bouche et le nez d'une nourriture avariée). D'autres émotions se sont développées au cours de l'évolution car elles ont permis une meilleure adaptation (*ex.* exprimer la peur permet de prévenir d'autres individus d'un danger). Enfin, certaines expressions d'émotions peuvent être exprimées délibérément, sans pour autant qu'une émotion soit ressentie (*ex.* pour illustrer le discours).

- *Expression d'attitudes :* Les attitudes interpersonnelles permettent d'exprimer le sentiment d'une personne envers une autre. Cette expression peut être spontanée ou contrôlée. Nous détaillons le concept d'attitude et comment celles-ci sont représentées et exprimées par le comportement dans la section 2.2.

- *Accompagner et supporter le discours :* Le comportement non-verbal permet de réguler la conversation et d'assurer le tour de parole. Par exemple, des

hochements de tête permettent d'indiquer à son interlocuteur qu'on l'écoute, et le regard permet de conserver ou de donner le tour de parole.

- *Présentation de soi :* Le comportement non-verbal et l'apparence (*ex.* les choix vestimentaires) permettent d'afficher sa personnalité ou son appartenance à certains groupes sociaux (*ex.* se montrer plus ou moins conformiste en choisissant de s'habiller de manière plus ou moins provocante dans des situations formelles, comme des diners ou des entretiens d'embauche).
- *Rituels :* Le comportement non-verbal permet d'assurer certains rituels sociaux, comme les salutations (*ex.* poignée de mains).

2.1.5 Gestes communicatifs

Les gestes sont définis comme les mouvements des bras et des mains non-manipulatifs ayant lieu au cours du discours [Kendon, 1983, McNeill, 1992].

Catégories de gestes :

Plusieurs classifications des types de gestes ont été proposées [Ekman & Friesen, 1969, McNeill, 1992]. McNeill présente les catégories de gestes suivantes :

- Les gestes *iconiques* illustrent une propriété concrète de l'objet ou de l'action dont on est en train de parler. Par exemple, pour indiquer qu'une personne est très grande, on peut lever le bras très haut avec la paume orientée vers le bas.
- Les gestes *métaphoriques* sont proches des gestes *iconiques* mais illustrent le discours d'une manière abstraite. On peut prendre l'exemple d'une personne disant « C'est une bonne histoire » en tenant un objet imaginaire entre ses deux mains : l'histoire est matérialisée par cet objet imaginaire.
- Les gestes *déictiques* sont des mouvements de pointage qui peuvent indiquer une personne, un objet, un lieu ou une direction concrète, mais aussi potentiellement des objets imaginaires.
- Les gestes *bâtons* (*beats*) sont des mouvements simples, rythmiques, qui accompagnent le discours, mais qui ne comprennent pas d'information sémantique.

Si les travaux d'Ekman et Friesen ne traitaient pas exclusivement des gestes, des liens peuvent être tracés entre les catégories des signaux *illustrateurs* (voir Section 2.1.1)

et les catégories de gestes de McNeill [McNeill, 1992]. Tout d'abord, on retrouve les gestes *bâtons* et *déictiques* dans les deux classifications. Les signaux *idéographiques* (« *movements which sketch a path of direction of thought* ») sont à rapprocher des gestes *métaphoriques*. Enfin, la catégorie des gestes *iconiques* peut rassembler à la fois les *mouvements spatiaux*, les signaux *kinétographiques* et *pictographiques*, dans le sens où ceux-ci illustrent un attribut concret (*resp.* une relation spatiale, un mouvement corporel, une image).

Le continuum de Kendon : relation entre le langage et les gestes :

Pour Kendon, les gestes font partie intégrante du langage [Kendon, 1988]. Celui-ci identifie différents types de gestes qu'il répartit sur un continuum selon leur relation avec le discours. Plus on se déplace vers la droite de ce continuum, plus le degré auquel le discours accompagne forcément le geste diminue, et plus les gestes suivent une convention et possèdent les propriétés d'un langage [McNeill, 2005] :

$$Gesticulation \rightarrow Speech - Framed\ Gestures \rightarrow Pantomime \rightarrow Emblems \rightarrow Signs$$

Les gestes de la catégorie *Gesticulation* sont co-verbaux, c'est à dire qu'ils accompagnent le discours (*ex.* bâtons). Leur sens peut être relié à une information présente dans le discours. Les *Speech-Framed Gestures* remplacent le langage parlé et constituent une partie du discours. Un exemple fourni par McNeill est le suivant : « *Les parents étaient sympathiques, mais les enfants étaient [geste]* ». Ici, le geste occupe la fonction grammaticale d'un adjectif [McNeill, 1992, p. 37]. Les *Pantomimes* consistent à mimer des actions et objets avec les mains. Les emblèmes sont des gestes communiquant un sens précis défini par convention (voir les travaux d'Ekman et Friesen, Section 2.1.1), et sont placés plus à droite sur le continuum que les *Pantomimes* car ceux-ci possèdent une forme standardisée. Par exemple, pour dire « OK » par le biais d'un geste, nous formons un cercle avec le pouce et l'index et nous l'orientons vers notre interlocuteur. Enfin, Kendon place les langages des signes à l'extrême droite de ce continuum, car ceux-ci possèdent les caractéristiques de systèmes linguistiques complets [McNeill, 1992].

Les *points de croissance* de McNeill :

McNeill avance que les gestes et le discours proviennent d'un même processus de pensée [McNeill, 2005]. Les gestes seraient une manifestation d'unités atomiques

de pensées impliquées dans la production du discours. Ces unités sont appelées les *points de croissance* (*Growth Points*). Un *point de croissance* n'est ni l'évocation d'un mot ou d'une image mais le fait de penser simultanément en termes d'imagerie globale et en catégories linguistiques (« *thinking in global imagery and linguistic categories simultaneously* »). De ces *points de croissance* proviennent la production de gestes et la production linguistique. McNeill déduit de cette origine commune la raison de la synchronisation entre gestes et parole.

Les différentes phases d'un geste :

Les gestes peuvent être décomposés en plusieurs phases [Kita et al., 1998] :

1. *Preparation (optionnelle) :* le bras se déplace d'une position de repos vers une position intermédiaire à partir de laquelle la partie du mouvement porteuse de sens (*stroke*) va être exécutée.
2. *Pre-stroke hold (optionnelle) :* la position obtenue après la phase de préparation peut être maintenue plus ou moins longtemps, par exemple pour synchroniser le discours avec le geste.
3. *Stroke (obligatoire) :* phase où le sens du geste est exprimé. Cette phase est synchronisée avec le segment du discours dont le geste partage le sens.
4. *Post-stroke hold (optionnelle) :* maintien optionnel de la position finale du geste pour synchroniser le geste et le discours.
5. *Retraction (optionnelle) :* les mains reviennent à la position de repos si la personne n'enchaine pas avec un autre geste.

2.1.6 Conclusion

Nous avons présenté des travaux fondateurs ayant étudié le comportement non-verbal, ses différentes catégories de signaux, et ses fonctions. Afin de tirer bénéfice de la capacité des Agents Conversationnels Animés à utiliser les mêmes modalités du comportement non-verbal que les humains, il est nécessaire de connaître les différentes fonctions que peuvent réaliser ces signaux non-verbaux. Par exemple, les gestes de l'Agent Conversationnel Animé vont accompagner le discours, et les types et les formes de ces gestes vont être déterminés par le contenu et la structure du

discours [McNeill, 1992, Cosnier & Vaysse, 1997, Poggi, 2003]. De la même manière, le comportement non-verbal possède une fonction régulatrice participant au bon déroulement d'une conversation (*ex.* indiquer à l'interlocuteur qu'il peut continuer à parler par le biais de hochements de tête) [Ekman & Friesen, 1969, Argyle, 1988, Cosnier & Vaysse, 1997], et un Agent Conversationnel Animé menant une conversation avec un utilisateur doit pouvoir assurer ces fonctions.

Dans la section suivante, nous nous intéressons en particulier aux attitudes sociales. Nous reprenons plusieurs définitions de ce concept et illustrons pourquoi les attitudes sont intéressantes dans un contexte de simulation aux entretiens d'embauche. Nous présentons ensuite la représentation des attitudes sociales que nous adoptons. Enfin, nous détaillons comment les signaux non-verbaux de différentes modalités contribuent à l'expression d'attitudes.

2.2 Attitudes : définitions et représentation

Dans cette section, nous évoquons plusieurs définitions de l'attitude sociale qui nous servent à illustrer les aspects importants de ce phénomène, et nous abordons la représentation que nous adoptons.

2.2.1 Définitions

Les attitudes (*attitudes* ou *stances* en anglais), sont un objet de recherche dans plusieurs domaines : en psychologie sociale, en linguistique sociale ou encore dans le traitement du signal social. Par conséquent, il existe de nombreuses définitions du concept d'attitude [Chindamo et al., 2012] . Nous retenons ici plusieurs définitions afin d'illustrer plusieurs aspects importants du concept d'attitude et d'expliquer la perspective que nous prenons pour guider nos travaux.

Dans le domaine de la linguistique sociale, Du Bois propose cette définition : « *Stance is a public act by a social actor, achieved dialogically through covert communicative means, of simultaneously evaluating objects, positioning subjects (self and others), and aligning with other subjects, with respect to any salient dimension of the sociocultural field.* » [Du Bois, 2007]. Ainsi, prendre ou exprimer une attitude par rapport à un objet, une situation ou quelqu'un, consisterait à l'*évaluer*, à se *positionner* par

rapport à lui, et à s'*aligner* ou non avec d'autres personnes. Dans notre contexte, nous nous intéressons à une attitude *interpersonnelle*, c'est à dire l'attitude que nous exprimons par rapport à quelqu'un d'autre : exprimer son attitude consiste ainsi à exprimer une évaluation et un positionnement par rapport à son interlocuteur.

L'expression d'une attitude est multimodale, comme le présente Chindamo : « *The expressive side of a stance includes unimodal as well as multimodal vocal or gestural (in a wide sense including all communicative and informative body movements) verbal or nonverbal contributions.* » [Chindamo et al., 2012]. Ainsi, l'attitude s'exprime à la fois par le biais du comportement verbal (*ex.* choix des mots et du niveau de langage) et non-verbal (*ex.* gestes, expressions faciales).

Scherer propose la définition suivante : « *The specificity of [Interpersonal stance] is that it is characteristic of an affective style that spontaneously develops or is strategically employed in the interaction with a person or a group of persons, coloring the interpersonal exchange in that situation (e.g. being polite, distant, cold, warm, supportive, contemptuous).* » [Scherer, 2005]. Un aspect intéressant de cette définition est que les attitudes sont soit spontanées, soit employées stratégiquement. Nous notons également que Scherer décrit l'attitude comme un « style affectif » qui « colore » une interaction. Ainsi, les attitudes ne sont pas un phénomène qui s'exprime à un instant précis ou de manière prototypique, mais plutôt tout au long d'une interaction ou au moins sur des intervalles de temps assez longs.

Exprimer une attitude interpersonnelle est donc un acte *spontané* ou *stratégique*, qui consiste ainsi à *évaluer* ou à se *positionner* socialement par rapport à son interlocuteur. Or, c'est effectivement un des objectifs d'un recruteur que d'arriver à évaluer socialement un candidat à un poste, par exemple pour s'assurer qu'il pourra s'intégrer dans une équipe. En outre, si un recruteur peut se comporter de manière naturelle (*i.e. spontanée*) par rapport à un candidat, il peut décider d'adapter *stratégiquement* l'attitude qu'il exprime afin d'évaluer les réactions sociales du candidat à une telle attitude. Ces deux aspects essentiels des attitudes sociales en font des phénomènes intéressants à reproduire par des recruteurs virtuels.

De plus, nous retenons de ces définitions que les attitudes sont exprimées par le biais du *comportement verbal* et du *comportement non-verbal*. Cependant, l'expression d'une attitude ne s'effectue pas à un instant donné (à l'inverse d'autres phénomènes, comme certaines émotions, qui s'expriment par une expression précise) mais sur de

plus longues périodes, en altérant le *style affectif* d'une personne.

Nous présentons à présent la représentation que nous adoptons pour les attitudes interpersonnelles.

2.2.2 Représentation

L'étude de la communication interpersonnelle a amené des chercheurs à essayer d'identifier des dimensions permettant de différencier et catégoriser différents types de comportements interpersonnels. Schutz propose les dimensions d'inclusion, de contrôle, et d'affection [Schutz, 1958]. Plus tard, Burgoon et Hale ont proposé jusqu'à 12 dimensions qui permettraient de définir et différencier plusieurs styles de communication : dominance, intimité, affection, intensité de l'engagement, inclusion, confiance, superficialité-profondeur, excitation émotionnelle, calme, similarité, formalité, orientation tâche ou sociale [Burgoon & Hale, 1984].

Argyle propose une représentation de l'attitude bi-dimensionnelle [Argyle, 1988]. Une première dimension est l'amicalité (*affiliation*), qui peut être caractérisée comme le fait de désirer une relation proche avec son interlocuteur ou non, et dont les valeurs positives représentent un comportement amical et les valeurs négatives représentent des comportements inamicaux ou hostiles. L'autre dimension est celle du statut (*status*), et représente la supériorité ou l'infériorité sociale d'une personne relativement à son interlocuteur. Ici, des valeurs positives représentent un comportement dominant, et inversement, des valeurs négatives représentent un comportement soumis. Ces deux dimensions peuvent être représentées sur un Circomplexe Interpersonnel (voir Figure 2.1) [Leary, 1957, Wiggins, 2003].

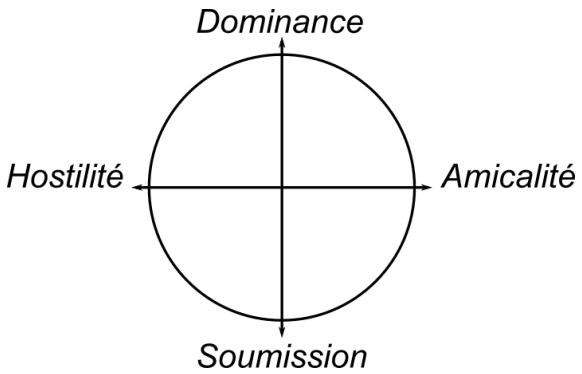


FIGURE 2.1: Circomplexe interpersonnel des dimensions de l'attitude.

Ces deux dimensions rejoignent les deux aspects étudiés par Mehrabian dans ses travaux sur le comportement non-verbal [Mehrabian, 1969] : celui-ci a étudié l'influence de la posture, de la distance, et du regard sur la communication du statut de la personne sur son interlocuteur (c'est à dire la dimension de *dominance* ou de *status*), et sur la communication de l'*affiliation* envers son interlocuteur (ce qu'il appelle pour sa part *attitude* : « *For the purposes of discussion, attitude is broadly defined as the degree of liking, positive evaluation, and/or preference of one person for another.* »).

Dans cette thèse, nous choisissons d'utiliser la représentation d'Argyle, comme de nombreux autres travaux dans le domaine des Agents Conversationnels Animés [Ballin et al., 2004, Lee & Marsella, 2011, Ravenet et al., 2013, Cafaro, 2014]. Cette représentation nous permet de nous appuyer sur les nombreux travaux ayant étudié le lien entre le comportement et l'expression de la dominance et de l'amicalité, afin d'identifier les modalités et signaux à considérer dans notre modèle. Nous détaillons à présent l'influence des différentes modalités du comportement sur l'attitude exprimée.

2.2.3 Expression de l'attitude

Plusieurs facteurs ont une influence sur l'attitude qui est exprimée par une personne. Dans nos travaux, nous nous concentrons sur l'influence du comportement non-verbal sur l'attitude perçue. Toutefois, nous présentons ici les indices associés à l'expression d'attitude pour toutes les modalités non-verbales, mais aussi vocales, verbales et pour l'apparence physique. Cet inventaire nous a permis d'identifier quelles modalités notre modèle de planification de comportement non-verbal devait considérer.

2.2.3.1 Apparence physique

L'apparence physique est source de nombreux préjugés et stéréotypes et influe sur l'attitude qui est attribuée à une personne, indépendamment d'un jugement fondé [Knapp et al., 2013]. Des personnes de grande taille ont tendance a être perçues comme plus dominantes, et des personnes plus attirantes ont tendance à être per-

çues comme étant d'un plus haut statut social et comme étant plus amicales [Vinciarelli et al., 2009]. Les personnes n'ayant pas l'air d'être en bonne santé ou ayant une apparence négligée sont moins appréciées, tandis que les personnes ayant une apparence saine, élégante ou propre sont jugées comme ayant plus d'estime de soi (ce qu'on peut rapprocher de la dimension de statut) [Naumann et al., 2009]. Enfin, la constitution physique (*somatotypes*, voir Figure 2.2) est aussi source de préjugés : les individus *endomorphes* sont vus comme plus sympathiques mais plus dépendant des autres, les *mésomorphes* sont vus comme plus confiants, matures et forts (dominants), et les *ectomorphes* sont vus comme plus nerveux et tendus (soumis) [Cortes & Gatti, 1965].

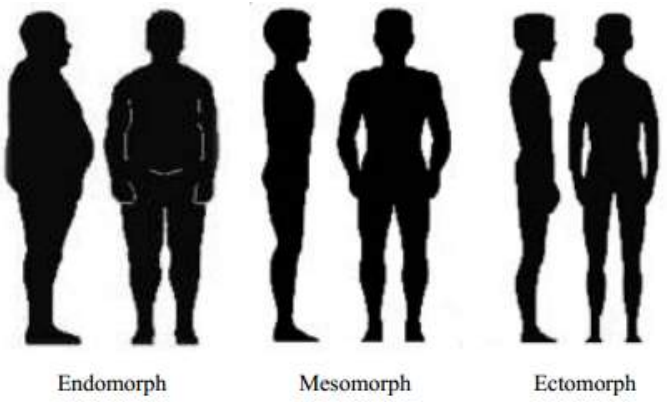


FIGURE 2.2: Somatotypes.

2.2.3.2 Comportement vocal et verbal

La prosodie (*i.e.* le rythme, la hauteur, l'amplitude de la voix) influence la manière dont une phrase est comprise, et sert notamment à indiquer qu'une phrase est une question (voix plus aigue à la fin de la phrase), à accentuer certaines parties du discours (voix plus forte et plus aigue sur les mots accentués) ou à le structurer (pauses, changements de rythme et d'intensité) [Hirschberg & Grosz, 1992, Hirschberg, 1993]. Mais la prosodie possède aussi une influence sur l'attitude qui est perçue : l'amplitude de la voix influence la perception de la dominance [Tusing & Dillard, 2000]. L'effet de la hauteur de la voix semble être différent selon le genre du locuteur : Scherer rapporte plusieurs études indiquant que des femmes plus dominantes présentaient une hauteur de voix plus basse, alors que le phénomène inverse est présent

pour les hommes, où les hommes à la voix la plus hautes sont plus dominants [Scherer & Scherer, 1981]. Enfin, un locuteur produisant peu d'hésitations est considéré comme plus dominant [Vinciarelli et al., 2009].

Les vocalisations linguistiques (*ex.* « hum », « ah ») sont aussi des signaux marquant des attitudes. Ainsi, des hésitations dans la voix peuvent indiquer de l'embarras [Glass et al., 1982]. Ces vocalisations peuvent aussi servir à produire des signaux en direction de l'interlocuteur lorsque l'on écoute, signaux appelés *backchannels* [Ward & Tsukahara, 2000]. Ceux-ci peuvent servir à exprimer l'attention, l'intérêt, mais aussi l'accord ou la contradiction [Allwood et al., 2007b, Vinciarelli et al., 2009]. Le rire est une vocalisation non-linguistique qui peut montrer de l'affiliation ou de l'amicalité [Keltner & Haidt, 1999].

Enfin, le comportement de prise de parole peut fortement exprimer de l'attitude : ainsi, interrompre son interlocuteur est vu comme un comportement hostile et dominant [Smith-Lovin & Brody, 1989]. Le degré de dominance d'une personne est directement lié à la quantité de temps où elle prend la parole [Gifford & Hine, 1994, Mast, 2002, Dunbar & Burgoon, 2005, Gatica-Perez, 2009].

2.2.3.3 Gestes

Si les gestes communicatifs accompagnent principalement le discours [McNeill, 1992], et sont utilisés pour communiquer certaines intentions (*ex.* lever le pouce pour féliciter ou montrer de l'appréciation) ou réguler le discours et l'interaction (*ex.* donner la parole, souligner un mot en levant un doigt, saluer), ils peuvent révéler des informations sur l'attitude d'une personne. Les gestes de la catégories des *adaptors* (*ex.* se gratter, manipuler un stylo) peuvent manifester un état de stress, et sont associés à de la soumission [Gifford, 1994, Knapp et al., 2013]. Toucher son interlocuteur est vu comme un signe d'amicalité et de dominance [Carney et al., 2005, Burgoon et al., 1984, Yabar & Hess, 2007].

L'expressivité des gestes (*ex.* intensité, amplitude, ouverture, ...) influe aussi sur l'attitude exprimée par une personne. Ainsi, une personne faisant des gestes larges est perçue comme plus dominante [Carney et al., 2005], et faire de nombreux gestes est aussi perçu comme un signe de dominance et d'amicalité [Gifford, 1994, Carney et al., 2005, Burgoon & Le Poire, 1999].

2.2.3.4 Posture et distance interpersonnelle

Les postures sont le plus souvent prises inconsciemment et sont ainsi un élément révélateur de l'état mental d'une personne [Richmond & McCroskey, 2000]. Adopter la même posture que son interlocuteur est aussi un signe d'amicalité [LaFrance, 1982]. De la même manière que des gestes prenant plus d'espace sont des signes de dominance, adopter une posture occupant un grand espace (*ex.* s'allonger dans sa chaise, mettre ses bras derrière la tête, *etc.*) est un signe de dominance. Inversement, une personne adoptant une posture plus recroquevillée est vue comme plus soumise [Mehrabian, 1977, Argyle, 1988, Gifford, 1994, Carney et al., 2005, Dunbar & Burgoon, 2005].

En outre, le positionnement spatial, en particulier la distance et l'orientation de son corps par rapport à son interlocuteur, participent eux aussi à l'expression d'attitudes sociales. Se positionner plus près de son interlocuteur ou plus orienté vers lui est un signe d'amicalité [Burgoon et al., 1984, Gifford, 1994, Burgoon & Le Poire, 1999, Yabar & Hess, 2007].

2.2.3.5 Regard

Les différentes fonctions du regard ont été étudiées dans le détail par Kendon, Argyle et Cook [Kendon, 1967, Argyle & Cook, 1976]. Si le regard est crucial dans la régulation de l'interaction (tour de parole, attention), il exprime aussi des attitudes [Argyle & Cook, 1976, Burgoon et al., 1984, Yabar & Hess, 2007, Briton & Hall, 1995]. Ainsi, une grande quantité de regards partagés (*mutual gaze*) indique une grande affiliation. De manière générale, regarder plus son interlocuteur est un signe d'amicalité, sauf si on regarde son interlocuteur quasiment en permanence, auquel cas le regard est alors considéré comme menaçant (hostile et dominant). Détourner le regard est perçu comme un signe de soumission, alors que ne pas détourner le regard est un signe de dominance.

2.2.3.6 Mouvements de tête

La tête est quasiment en permanence en mouvement lorsque nous parlons [Hadar et al., 1983]. Ses mouvements, ses positions, et ses combinaisons avec des signaux d'autres modalités donnent de nombreuses informations aux significations multiples [Heylen, 2008]. Cependant, la tête est paradoxalement une des modalités les moins étudiées du domaine de l'étude du comportement non-verbal [Heylen, 2008]. Ses fonctions communicatives accompagnant le discours sont en tout cas bien étudiées : les mouvements de tête accompagnent la prosodie, en particulier les accents sur certain mots, et peuvent signaler à l'interlocuteur que l'on est en train de penser [Munhall et al., 2004, McClave, 2000]. Les hochements de tête servent aussi à signaler l'accord, le désaccord (selon qu'ils sont verticaux ou horizontaux) ou la compréhension [Heylen, 2008].

Mais la tête est aussi une modalité impliquée dans l'expression des attitudes. Ainsi, l'angle d'inclinaison la tête indique de la dominance quand la tête est orientée vers le haut (garder « la tête haute ») et de la soumission quand elle est penchée vers le bas [Mignault & Chaudhuri, 2003, Gifford, 1991]. Pencher la tête sur le côté (*head canting* ou *cocking*) est par contre un signe d'amicalité [Otta et al., 1994, Debras & Cienki, 2012] ou de soumission [Otta et al., 1994, Poggi & Pelachaud, 1998, Costa et al., 2001]. Les hochements de tête horizontaux (*head shakes*) sont typiquement associés avec l'expression de la négativité ou du désaccord, et sont observés plus fréquemment chez des personnes dominantes [Gifford, 1994, Carney et al., 2005]. Pour les hochements de tête verticaux, ils semblent être plus présents chez des personnes amicales [Burgoon & Le Poire, 1999, Harrigan et al., 1985, Gifford, 1991], mais il est plus difficile d'obtenir un consensus dans la littérature sur leur lien avec la dimension de dominance ou de soumission : certaines études ne rapportent pas d'effet des hochements de tête sur les attitudes exprimées [Harrigan et al., 1985, Gifford, 1991], tandis que d'autres études rapportent une corrélation (faible) de la quantité de hochements de tête verticaux avec la dominance [Burgoon & Le Poire, 1999], ou inversement une corrélation avec la soumission dans le cas d'une étude utilisant des agents virtuels [Von der Puetten et al., 2010].

2.2.3.7 Expressions faciales

Les expressions faciales ont été largement étudiées dans le cadre de l'expression des émotions [Darwin, 1872, Ekman & Friesen, 1969]. La propension d'un individu à afficher certaines émotions ou à les inhiber a quant à elle été reliée avec la perception de l'attitude de cet individu [Knutson, 1996, Carney et al., 2005] : les expressions de joie sont associées à l'amicalité et la dominance, la colère et le dégoût avec l'hostilité et la dominance, tandis que la peur et la tristesse sont associées avec la soumission.

Les mouvements du visage ont été aussi été étudiés indépendamment des émotions : ainsi, les sourires sont typiquement associés à des attitudes amicales et de soumission [Keating et al., 1981, Otta et al., 1994, Knutson, 1996, Hess et al., 2000, Dunbar & Burgoon, 2005, Krumhuber et al., 2007]. L'effet des froncements de sourcils est quant à lui dépendant de la culture des personnes impliquées : dans des cultures occidentales, ils correspondent à des personnes plus dominantes [Aronoff et al., 1988, Keating et al., 1981], tandis que Keating *et al.* observent qu'en Thaïlande rurale, ce sont les haussements de sourcils qui sont vu comme plus dominants [Keating et al., 1981].

Dans la prochaine section, nous présentons les perspectives d'interprétation du comportement non-verbal qui peuvent influencer la perception de celui-ci, et qu'il faut donc considérer dans la planification du comportement non-verbal.

2.3 Les différentes perspectives d'interprétation des signaux non-verbaux

Argyle décrit la communication non-verbale de la manière suivante [Argyle, 1988]. « *Non-verbal communication [...] takes place whenever one person influences another by means of facial expression, tone of voice, or any of the others channels [of non-verbal communication]. This may be intentional, or it may not be.* ». Argyle propose un modèle de la communication, où une personne A, l'*émetteur*, communique une information à son interlocuteur B, le *récepteur*, par le biais d'un signal s intentionnel ou non. Ce modèle est représenté par la Figure 2.3. Selon celui-ci, il suffit de connaitre comment un signal non-verbal est interprété (*i.e. décodé*) afin

de planifier le comportement non-verbal d'un Agent Conversationnel Animé. Par exemple, si on veut que l'ACA exprime une attitude amicale, il suffira de choisir un signal non-verbal à exprimer qui est décodé comme exprimant une attitude amicale. Cependant, l'interprétation d'un signal non-verbal peut se faire sous différentes perspectives que nous présentons ci-dessous.

FIGURE 2.3: Modèle de la communication d'Argyle [Argyle, 1988]. Notons que les processus d'*encodage* et de *décodage* ne sont pas symétriques : il n'est pas garanti qu'une information envoyée par A par le biais d'un signal s soit correctement interprétée par B.

1. Certains signaux non-verbaux sont porteurs d'un sens précis. Par exemple, les gestes *emblèmes* représentent des concepts de manière conventionnelle dans une culture donnée. Ainsi, dans la culture occidentale, un pouce vers le haut indique que l'on félicite quelqu'un ou qu'on approuve, tandis qu'orienter sa paume verticalement vers quelqu'un lui indique de s'arrêter [Poggi, 2007]. De plus, certains signaux étudiés de manière isolée peuvent indiquer des états mentaux ou des attitudes : un sourire est ainsi vu généralement comme un signe d'amicalité [Burgoon et al., 1984].

2. Mais la présence d'autres signaux non-verbaux simultanés ou proches peut influencer le sens d'un signal. Comme l'indiquent Burgoon et Le Poire [Burgoon & Le Poire, 1999], on ne peut pas analyser un signal non-verbal de manière isolée : « *What illuminates the interpretation of a given behavior is its accompanying composite of nonverbal cues. No nonverbal cue is an "island". It is continually surrounded by a host of nonverbal behaviors which together may delimit and clarify meaning.* ». Pour interpréter des signaux non-verbaux il faut ainsi prendre en compte les signaux multimodaux proches. Ainsi, Keltner montre que c'est par l'enchainement temporel précis d'un sourire, d'un évitement de regard et d'un détournement de tête que l'on peut différencier l'embarras, l'amusement et la honte [Keltner, 1995]. Récemment, d'autres auteurs ont apporté un nouveau regard sur l'analyse des expressions faciales des émotions, typiquement analysées à partir de photographies (vues statiques de face du visage d'un acteur ou d'une personne exprimant une émotion), en

utilisant des vidéos et des techniques d'analyse séquentielle [With & Kaiser, 2011, Jack et al., 2014]. Ils ont démontré que l'enchainement des signaux faciaux et leur combinaison temporelle, spatiale et multimodale avec d'autres signaux non-verbaux jouent un rôle important et encore sous-étudié dans la perception des émotions.

3. De plus, la communication ne s'effectue pas dans un seul sens. Les personnes en interaction produisent aussi des signaux en réaction à ceux de leur(s) interlocuteur(s). Une personne en train d'écouter communique régulièrement à la personne en train de parler certains de ses états mentaux, comme son degré de compréhension ou d'intérêt, par le biais de signaux subtils, comme des hochements de tête ou vocalisations [Allwood et al., 1992, Poggi, 2007]. Ces signaux, appelés « *backchannels* », sont émis par la personne en train d'écouter à des moments précis du discours de son interlocuteur : après des pauses d'une durée précise, par exemple [Ward & Tsukahara, 2000]. La tendance à imiter ou non les signaux de son interlocuteur, par exemple à adopter la même posture que lui (*mimicry* ou imitation), est aussi révélatrice de son attitude envers son interlocuteur [LaFrance, 1982].

4. Enfin, certaines informations peuvent être obtenues en observant le comportement non-verbal d'une personne sur de plus longues périodes. Par exemple, une personne regardant son interlocuteur dans les yeux est vue comme digne de confiance [Mason et al., 2005], jusqu'à un certain point. En effet, fixer trop longtemps son interlocuteur du regard est vu comme agressif [Argyle & Cook, 1976]. Une personne produisant beaucoup de gestes est considérée comme plus dominante [Dunbar & Burgoon, 2005, Burgoon & Dunbar, 2006].

Une première perspective d'interprétation d'un signal non-verbal correspond au modèle d'Argyle : un signal s est envoyé par A et décodé par B, sans autre information. Nous appelons cela la perspective *unimodale*. Dans le deuxième cas, B interprète le dernier signal s envoyé par A dans le contexte d'une séquence de signaux $< s_1, ..., s_{i-1}, s >$. C'est la perspective *séquentielle*. Dans le troisième cas, B interprète le signal s de A par rapport à un signal qu'il a lui même envoyé, s_B. C'est la perspective *interactionnelle*. Enfin, dans le dernier cas, B interprète les tendances de A à émettre des signaux non-verbaux sur de longues périodes, tendances que nous notons S. C'est la perspective d'interprétation *globale*. Nous étendons ainsi le

modèle de la communication d'Argyle pour intégrer ces différentes perspectives (voir Figure 2.4).

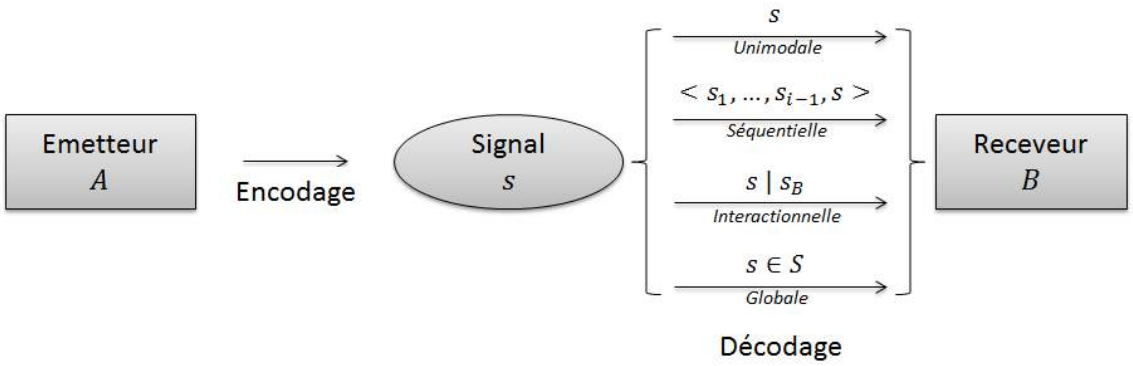


FIGURE 2.4: Modèle de la communication d'Argyle revisité avec les quatre perspectives d'interprétation que nous avons identifiées.

Voici un exemple de la manière dont ces différentes perspectives permettent d'analyser le comportement non-verbal de personnes en interaction. Imaginez un recruteur qui pense que le candidat avec lequel il est en train de s'entretenir a menti en indiquant maitriser l'anglais sur son CV, et qu'il veut le tester sur ce sujet. Le recruteur étend sa main droite vers le candidat tout en posant la question « Vous prétendez parler anglais couramment. Pouvez-vous me le prouver ? ». Le candidat regarde vers le côté pendant un instant, réfléchissant et hésitant. Il oriente enfin son regard vers le recruteur et essaie une réponse, puis esquisse un léger sourire suivi d'un évitement de regard et d'un détournement de la tête. Pendant que le candidat est en train de parler, le recruteur fronce les sourcils, et hoche la tête horizontalement alors que le candidat termine. Durant tout ce temps, le recruteur n'a de cesse de regarder le candidat dans les yeux.

Dans cet exemple, le geste réalisé par le recruteur est utilisé pour indiquer qu'il pose une question au candidat et qu'il lui donne la parole afin qu'il y réponde. Ce geste remplit ainsi une fonction communicative liée à la régulation de la conversation. En répondant, le candidat sourit, cependant ce n'est pas un sourire d'amicalité mais de soumission car le sens de ce sourire est altéré par les signaux multimodaux suivants (évitement de regard et détournement de tête). Le langage corporel du recruteur pendant la réponse du candidat, le froncement de sourcil et le hochement de tête horizontal, indiquent sa réaction de désaccord, et son inamicalité. Enfin, que le recruteur ait fixé des yeux le candidat pendant une longue durée de temps est un

signe de dominance.

Plusieurs informations peuvent être tirées de ces quatre perspectives d'interprétation des signaux non-verbaux. Notre point de vue est que celles-ci doivent toutes être considérées afin de pouvoir planifier correctement le comportement d'un Agent Conversationnel Animé. Comme nous l'avons montré en Section 2.2.3, de nombreux travaux existant nous fournissent des informations sur l'attitude exprimée par des signaux vus indépendamment (perspective unimodale), et d'autres travaux nous donnent aussi des informations sur les attitudes exprimées par des signaux en réaction à ceux de l'interlocuteur (perspective interactionnelle), ou dans les tendances globales du comportement (perspective globale). Cependant, les attitudes exprimées par des séquences de signaux non-verbaux n'ont pas été étudiées de manière systématique.

Conclusion

Afin de pouvoir planifier le comportement non-verbal d'Agents Conversationnels Animés, il est nécessaire de connaître les différentes fonctions du comportement non-verbal, et le lien entre les signaux des différentes modalités non-verbales et ces fonctions. Nous avons ainsi commencé par présenter des travaux fondateurs dans le domaine de l'analyse et la caractérisation du comportement non-verbal. Le comportement non-verbal peut participer à la régulation de l'interaction et de la conversation (signaux *régulateurs* d'Ekman et Friesen, *synchronisateurs* de Cosnier et Vaysse). Il peut servir à illustrer (gestes *illustrateurs*, *déictiques*, *métaphoriques*, *iconiques*, signaux *référentiels*), rythmer le discours (*bâtons*, signaux *paraverbaux*), voire même à le remplacer (*emblèmes*). De plus, il permet d'exprimer des émotions ou des attitudes sociales (*manifestations affectives*, signaux *expressifs*).

Nous nous sommes intéressés particulièrement à l'expression des attitudes sociales. L'attitude sociale permet de se *positionner* socialement par rapport à un interlocuteur : on peut se montrer plus ou moins amical, ou plus ou moins dominant par rapport à celui-ci. C'est un concept intéressant car l'attitude sociale peut être adoptée *spontanément* ou *stratégiquement*. Or, un recruteur peut, lors de situations d'entretiens d'embauches, décider d'adopter une attitude particulière afin de tester les réactions d'un candidat. Ainsi, la simulation des attitudes sociales est nécessaire

pour l'entrainement aux entretiens d'embauches car les candidats doivent pouvoir être confrontés aux différentes attitudes qui pourraient être adoptées par des recruteurs réels. Nous avons étudié et détaillé le lien entre les signaux des différentes modalités du comportement non-verbal et l'expression de l'attitude.

Enfin, nous avons présenté le modèle de la communication d'Argyle. Dans ce modèle, un émetteur *encode* des messages par le biais de signaux non-verbaux (intentionnellement ou non). Un receveur reçoit ces signaux, et les interprète (c'est le *décodage*). Ce modèle ne considère pas qu'un signal non-verbal peut être interprété selon différentes perspectives : en effet, des travaux montrent qu'un signal non-verbal peut aussi être interprété en tant que partie d'une séquence de signaux non-verbaux, en tant que réaction à un signal d'un interlocuteur, et par rapport aux tendances de comportement globales d'une personne. Nous avançons qu'un système de planification de comportement non-verbal devrait considérer ces quatre perspectives d'interprétation des signaux non-verbaux.

Cela est particulièrement le cas pour un modèle de planification de comportement pour l'expression d'attitudes. En effet, la définition de Scherer [Scherer, 2005] nous indique que les attitudes ne s'expriment pas à des instants donnés, mais tout au long d'une interaction. Ainsi, les signaux non-verbaux choisis lorsqu'un ACA prend la parole ne pourront pas être calculés indépendamment les uns des autres, mais devront être considérés ensemble afin d'assurer une cohérence dans l'attitude qu'ils expriment. A cette fin, nous proposons dans notre thèse de planifier le comportement d'ACA exprimant des attitudes par le biais de séquences de signaux non-verbaux.

Synthèse du chapitre

1. Les signaux non-verbaux remplissent plusieurs fonctions dans l'interaction. Ils peuvent participer à la régulation de l'interaction, compléter et structurer le discours. Enfin, ils servent à l'expression d'émotions et d'attitudes.
2. De nombreuses définitions ont été proposées pour définir les attitude sociales. L'attitude sociale consiste à se *positionner socialement* par rapport à une personne. C'est un phénomène *spontané* ou bien adopté *stratégiquement*, et qui s'exprime par le comportement *multimodal*. Nous adoptons la représentation d'Argyle, comportant une dimension d'*amicalité* et une dimension de *dominance*. Des signaux de nombreuses modalités participent à l'expression de l'attitude sociale.
3. L'interprétation d'un signal non-verbal ne se fait pas toujours de manière indépendante. Le sens d'un signal peut être influencé par les signaux multimodaux proches, les signaux de l'interlocuteur, ou les tendances globales de comportement de la personne.

3

État de l'art

Notre travail de recherche se situe dans le contexte des applications d'entrainement social utilisant des Agents Conversationnels Animés. Plusieurs applications de ce type ont d'ores et déjà été proposées. Elles présentent en effet plusieurs intérêts majeurs. Elles permettent un entrainement répété, alors que l'entrainement avec des partenaires ou des spécialistes est contraint par les disponibilités des autres personnes. De plus, ces applications permettent un entrainement standardisé, alors que l'entrainement avec d'autres personnes est forcément altéré par des aspects socio-émotionnels comme la relation avec ces personnes (*ex.* un ami avec lequel s'entrainer n'est peut-être pas objectif vis à vis de notre performance). Enfin, l'entrainement avec des Agents Conversationnels Animés permettrait de réduire certaines appréhensions sociales comme la peur d'être jugé ou de se sentir ridicule.

Si les applications d'entrainement social utilisant des ACAs présentent des avantages indéniables, il est cependant essentiel de s'assurer que ces applications améliorent effectivement les compétences sociales des utilisateurs. En outre, il est nécessaire d'identifier les conditions permettant d'obtenir les meilleurs résultats d'apprentissage lors d'utilisations de ces applications. Un autre enjeu majeur dans la conception d'ACAs pour l'entrainement social est la simulation des différents comportements socio-émotionnels auxquels un utilisateur doit pouvoir être confronté. Par exemple, des recruteurs peuvent exprimer différentes attitudes sociales envers un candidat lors d'entretiens d'embauches réels. Des ACAs utilisés pour entrainer aux entretiens d'embauche doivent donc pouvoir exprimer de telles attitudes.

Comme nous l'avons montré au chapitre 2, les attitudes s'expriment en particulier par le biais des signaux non-verbaux. Nous avons observé que le sens de certains signaux non-verbaux peut être influencé par la séquence de signaux dans laquelle ils se trouvent. De plus, les attitudes ne s'expriment pas à un instant donné mais au cours de longues périodes. Lors de la planification de signaux non-verbaux pour

l'expression d'attitudes, il faut donc considérer ces signaux ensemble et non pas indépendamment les uns des autres afin de s'assurer de la cohérence de l'attitude exprimée.

Dans ce chapitre, nous présentons une revue de la littérature du domaine des Agents Conversationnels Animés selon trois thématiques. Nous nous sommes intéressés aux systèmes existants pour l'entrainement social utilisant des ACAs, en particulier pour l'entrainement aux entretiens d'embauche et pour l'entrainement de la compétence de prise de parole en public. Nous présentons ces approches et leurs limites dans la section 3.1. Dans le chapitre 2, nous avons présenté les attitudes sociales et pourquoi des recruteurs virtuels devraient pouvoir en exprimer. Des modèles computationnels ont été proposés pour l'expression d'attitudes par des ACAs : nous les présentons dans la section 3.2. Enfin, puisque le sens d'un signal non-verbal peut être influencé par la présence d'autres signaux non-verbaux proches (voir Section 2.3), nous avons étudié les modèles de planification de comportement d'ACAs prenant en compte des séquences de signaux non-verbaux, et nous les détaillons dans la section 3.3.

3.1 Agents Conversationnels Animés pour l'entrainement social

Les systèmes d'entrainement de compétences sociales utilisant des personnages virtuels présentent plusieurs avantages par rapport aux approches traditionnelles, que ce soit avec mise en situation (*ex.* jeux de rôle) ou sans (*ex.* formations par cours magistraux) [Hart et al., 2013, Swartout et al., 2013]. A l'inverse d'approches traditionnelles utilisant des mises en situation, les systèmes utilisant des personnages virtuels ne sont pas exposés aux problèmes de disponibilité des intervenants, d'organisation, et de coûts. De plus, les jeux de rôle et les mises en situation avec des acteurs et tuteurs humains sont soumis à des variables sur lesquels ces intervenants n'ont pas le contrôle (*ex.* leur culture, leur genre, leur langue, leur âge, *etc.*). Inversement, l'apparence et les comportements de personnages virtuels peuvent être programmés, contrôlés et présentés précisément et de manière systématique. Le rythme et le type des situations d'entrainement peuvent être contrôlés et standardisés. On peut par exemple commencer l'entrainement à un niveau relativement facile, afin de motiver l'utilisateur, et accentuer progressivement la difficulté et la variété des situations présentées. Un autre avantage est que la disponibilité, l'accessibilité et la portée

des systèmes d'entrainement social utilisant des Agents Conversationnels Animés ne sont limitées que par l'accès à ces systèmes.

Des résultats récents suggèrent que l'interaction avec des personnages virtuels peut réduire la peur associée à la perception d'être jugé par autrui [Gratch et al., 2014], et ainsi réduire l'appréhension de personnes à s'entrainer [Hart et al., 2013]. De la même manière, de tels outils pourraient aider les personnes socialement anxieuses à réduire leur anxiété par la pratique dans un environnement sûr [Harris et al., 2002, Grillon et al., 2006]. Enfin, des travaux ont montré que les Agents Conversationnels Animés peuvent réussir à renforcer l'engagement ou la motivation de l'utilisateur [Gratch et al., 2006]. Or, ces facteurs peuvent influencer favorablement un processus d'apprentissage [Moreno et al., 2001, Rowe et al., 2010, Hart et al., 2013].

Récemment, de nombreuses applications d'entrainement social utilisant des Agents Conversationnels Animés ont été proposées. L'application *FearNot !* développée dans le projet européen FP6 eCIRCUS avait pour but de fournir à des enfants des stratégies pour faire face au harcèlement [Aylett et al., 2007]. Une reproduction du jeu populaire du loup-garou a été réalisée afin de sensibiliser les enfants aux différences culturelles [Aylett et al., 2014]. Des approches ont aussi été proposées pour aider les enfants diagnostiqués avec des troubles autistiques dans l'acquisition de compétences sociales [Bosseler & Massaro, 2003, Bernardini et al., 2012]. D'autres systèmes ont été proposés pour l'amélioration des compétences sociales de personnels de santé, en simulant par exemple des patients virtuels [Talbot et al., 2012, Johnsen et al., 2007, Rizzo et al., 2011]. Une autre thématique est celle de l'entrainement interculturel [Johnson & Valente, 2009, Ogan & Lane, 2010, Lane et al., 2013]. Par exemple, le système BiLAT (voir Figure 3.1) a été développé pour l'entrainement de soldats américains à préparer et mener des entretiens et négociations avec des interlocuteurs de culture irakienne [Lane et al., 2013].

Dans la section suivante, nous présentons plus particulièrement un état de l'art sur les systèmes simulant des audiences virtuelles pour la simulation de prises de parole en public, l'entrainement de la compétence de prise de parole en public étant un des domaines d'application de notre thèse (voir Chapitre 8).

FIGURE 3.1: Capture d'écran du système BiLAT [Lane et al., 2013].

3.1.1 Simulation de prise de parole en public

Des audiences virtuelles ont été étudiées pour leur potentiel à réduire l'anxiété liée aux situations de prise de parole en public [Pertaub et al., 2002, Harris et al., 2002, Grillon et al., 2006]. Les résultats de ces études suggèrent que la pratique de la prise de parole en public devant une audience virtuelle permettrait de réduire les niveaux d'anxiété des participants. Nous présentons plus précisément ces études ci-dessous.

Harris *et al.* ont étudié l'utilisation d'une audience virtuelle pour la réduction de l'anxiété, avec une population d'étudiants présentant habituellement une forte anxiété lors de prises de parole en public [Harris et al., 2002]. La moitié des participants étaient exposés à un système de réalité virtuelle simulant une situation de prise de parole en public, une fois par semaine pendant quatre semaines. L'autre moitié des participants était assignée à une liste d'attente (groupe de contrôle). Les participants du premier groupe utilisaient un casque de réalité virtuelle dans lequel était affiché un environnement 3D reproduisant un amphithéâtre d'université, pouvant accueillir jusqu'à cent personnages virtuels. Lors de la première séance, l'amphithéâtre était

vide. Dans la deuxième séance, celui-ci se remplissait progressivement de personnages virtuels, qui encourageraient et applaudissaient le participant. Dans les deux dernières séances, les personnages présents dans l'amphithéâtre demandaient au participant de parler plus fort, riaient ou parlaient entre eux (*i.e.* ils ne portaient pas leur attention au participant). Le niveau d'anxiété des participants était mesuré par des questionnaires et par une mesure du rythme cardiaque. Les auteurs observent des différences significatives entre le groupe assigné au système et le groupe de contrôle sur l'ensemble des mesures. Ces résultats suggèrent que des sessions d'interaction répétées avec des audiences virtuelles sont efficaces pour réduire l'anxiété à la prise de parole.

Grillon *et al.* ont eux aussi étudié l'impact d'expositions répétées à un système de réalité virtuelle [Grillon et al., 2006]. De la même manière que dans l'étude présentée ci-dessus [Harris et al., 2002], un casque de réalité virtuelle était utilisé. Cependant, le système utilisé (développé par Herbelin dans son travail de thèse [Herbelin, 2005]) permettait de confronter les participants à une plus grande variété de situations. Ainsi, les participants étaient d'abord exposés à un seul personnage virtuel dans un bureau, puis à un groupe de personnes, puis à un auditoire dans un amphithéâtre, puis à un groupe de personnes dans une cafétéria, *etc.* Les personnages virtuels disposaient d'un ensemble d'animations générales (*ex.* passer la main dans les cheveux, croiser les jambes), et de la capacité de modifier leur direction de regard et d'afficher quelques expressions faciales. Cependant, ces comportements n'étaient pas automatiques mais sélectionnés par la personne manipulant l'expérience. Le niveau d'anxiété était mesuré par des questionnaires, et la direction du regard des participants était enregistrée lors des interactions. Les résultats des questionnaires montrent une réduction de l'anxiété des participants. De plus, les données de suivi de regard indiquent que les participants focalisent plus leur regard sur les personnages virtuels après plusieurs séances de réalité virtuelle, ce que les auteurs interprètent comme un effet de la réduction de leur anxiété.

Enfin, Pertaub *et al.* ont étudié l'effet de différents types d'audiences virtuelles et du niveau d'immersion sur l'anxiété de participants [Pertaub et al., 2002]. Ils comparent trois audiences composées de huit personnages. Une des trois audiences est neutre (statique), la deuxième se comporte de manière positive (expressions de joie, hochements de têtes) et la dernière se comporte de manière négative (bâillements, endormissements, émotions négatives, voire sorties de salle). En outre, des phrases

peuvent être prononcées par l'audience (*ex. « C'est intéressant »* pour l'audience positive, *« Je vois »* pour la neutre, ou *« C'est vraiment n'importe quoi »* pour la négative). Ces phrases sont contrôlées par un opérateur (magicien d'Oz) qui choisit l'instant où elles se déclenchent. La moitié des participants était équipée d'un casque de réalité virtuelle, tandis que les autres participants voyaient l'audience par un simple écran de bureau. Le niveau d'anxiété des participants était mesuré par des questionnaires (*Personal Report of Confidence as a Public Speaker* [Paul, 1966]). Les résultats de cette étude montrent que l'audience négative parvient à provoquer de l'anxiété chez les participants, indépendamment de leur niveau habituel d'anxiété dans des situations de prise de parole en public (*i.e.* même des participants ayant confiance en eux déclarent ressentir de l'anxiété lors de la présentation devant l'audience négative). Les participants utilisant le casque de réalité virtuelle ressentent aussi une plus grande anxiété, en particulier les participants de sexe féminin (*i.e.* les femmes interagissant avec l'audience par le biais du casque de réalité virtuelle ressentent significativement plus d'anxiété que les hommes).

FIGURE 3.2: Audiences du système de [Pertaub et al., 2002]. L'audience adopte un comportement positif à gauche et négatif à droite.

La limite principale des différentes audiences virtuelles que nous avons présentées est qu'elles ne sont pas interactives. Si les personnages pouvaient exprimer certains signaux de diverses natures, ceux-ci n'étaient pas liés directement au comportement du participant, ou à sa performance. De plus, l'objectif des études présentées n'était pas d'améliorer directement la compétence de prise de parole en public des participants, mais d'étudier la réduction de l'anxiété liée à la prise de parole en public. Dans ce travail de thèse, nous avons développé une audience virtuelle interactive pour améliorer les compétences de prise de parole en public d'utilisateurs. Cette audience est détaillée dans le chapitre 8.

Une deuxième application d'entrainement social à laquelle nous nous sommes in-

téressés dans ce travail de recherche est l'entrainement aux entretiens d'embauche. Dans la section suivante, nous présentons un état de l'art de systèmes permettant de simuler des entretiens d'embauche.

3.1.2 Simulation d'entretiens d'embauche

D'autres auteurs ont étudié des systèmes de simulation d'entretiens d'embauche virtuels [Brundage et al., 2006, Kwon et al., 2013, Hoque et al., 2013].

Brundage *et al.* ont étudié l'influence du comportement d'un recruteur virtuel sur la quantité de bégaiements de personnes souffrant d'anxiété sociale [Brundage et al., 2006]. Deux recruteurs différents étaient étudiés, un plutôt encourageant et l'autre plus provocant : les différences dans le comportement de ces recruteurs se trouvaient en particulier dans la probabilité d'interrompre le participant et dans la quantité de regards dirigés vers le participant. Les auteurs observent que la quantité de bégaiements des participants lors d'entretiens d'embauches virtuels est corrélée avec celle observée lors d'entretiens réels. Cela suggère que les entretiens d'embauche virtuels génèrent de l'anxiété de la même manière que des entretiens réels. De plus, le type de comportement verbal et non-verbal du recruteur influe de manière significative sur la quantité de bégaiements du candidat.

Une étude sur l'influence du réalisme visuel sur l'anxiété engendrée par une simulation d'entretien d'embauche a été réalisée par Kwon *et al.* [Kwon et al., 2013]. Le recruteur était décliné en quatre versions : une photographie d'un recruteur réel, un personnage 3D non-réaliste (*cartoon*), un personnage 3D réaliste, et un recruteur réel. Les auteurs observent que la situation la plus stressante est l'entretien avec le recruteur réel, mais le personnage 3D réaliste parvient quand même à induire de l'anxiété à un niveau proche de celui du recruteur réel. De même, le personnage 3D non-réaliste parvient à créer un niveau d'anxiété significatif chez l'utilisateur. Le comportement des recruteurs virtuels était défini à l'avance (scripté) et ne s'adaptait pas à l'utilisateur.

Enfin, le système d'entrainement aux entretiens d'embauche *My Automated Conversation Coach* [Hoque et al., 2013] (MACH) fait figure d'exception, car il a pour but d'améliorer directement le comportement non-verbal des participants, en le mesurant automatiquement pendant un entretien d'embauche virtuel, puis en faisant un

rapport à la fin de l'interaction. MACH a été testé avec 90 étudiants de licence (*undergraduate students*) du campus du Massachussets Institute of Technology. L'étude consistait en trois phases où les participants interagissaient en premier lieu avec un conseiller humain, interaction qui servait à juger le niveau initial de compétences des participants (*pre-test*), puis, selon la condition qui leur était assignée, interagissaient avec une version spécifique de MACH (avec ou sans rapport à la fin de l'entrainement) ou regardaient simplement une vidéo de 30 minutes dédiée à la préparation aux entretiens d'embauche. Finalement, un entretien était réalisé avec le même conseiller humain qu'avant l'entrainement, afin d'évaluer le niveau d'amélioration (ou l'absence d'amélioration) suivant l'apprentissage. Les conseillers humains ne connaissaient pas la condition d'apprentissage des apprenants. Les résultats ont montré une amélioration significative des compétences des participants entre la condition utilisant le système MACH avec rapport final et la condition sans rapport final, et la condition utilisant le système MACH avec rapport final comparé à la condition où les participants regardaient une vidéo. La limitation principale de MACH est que le comportement du recruteur pendant l'interaction n'est pas affecté par la performance de l'utilisateur. C'est uniquement après l'interaction qu'un bilan sur sa performance lui est présenté. De plus, ce système ne permet pas de se confronter à plusieurs types de recruteurs, hormis la possibilité de choisir un recruteur de sexe masculin ou féminin.

FIGURE 3.3: Utilisateur interagissant avec le système MACH [Hoque et al., 2013].

A l'exception de MACH [Hoque et al., 2013], la plupart des systèmes d'audiences virtuelles ou d'entretiens d'embauche virtuels que nous avons présentés étaient fo-

calisés sur la réduction de l'anxiété du participant, et pas sur l'amélioration de ses compétences. Ils n'ont en tout cas pas été évalués en ce sens, et il est difficile de conclure si ces systèmes peuvent participer à l'amélioration des compétences sociales d'un utilisateur. En revanche, une évaluation du système MACH a montré que celui-ci permet d'améliorer le comportement des utilisateurs lors d'entretiens d'embauche [Hoque et al., 2013].

De plus, tous les systèmes que nous avons évoqués présentent une limitation centrale : même dans les cas où l'agent utilisé est interactif (*ex.* le recruteur virtuel de MACH pouvait produire des *backchannels*), le comportement de celui-ci n'est pas affecté par le comportement ou la performance de l'utilisateur. Nous notons également qu'aucun de ces systèmes de simulation d'entretiens d'embauche ne modélise les aspects socio-émotionnels d'une telle situation. Or, un entretien d'embauche est une situation sociale où le recruteur peut exprimer différentes attitudes envers le candidat. Certains recruteurs seront plus ou moins amicaux, plus ou moins dominants. Il est ainsi nécessaire qu'un système d'entrainement aux entretiens d'embauche puisse préparer des apprenants à de telles situations par le biais d'un recruteur capable d'exprimer différentes attitudes.

3.2 Modèles computationnels d'expression d'attitudes sociales

L'un des objectifs des travaux de recherche menés dans cette thèse est de doter un Agent Conversationnel Animé utilisé pour l'entrainement aux entretiens d'embauches de la capacité d'exprimer des attitudes sociales. Dans cette perspective, nous détaillons dans cette section des travaux existants sur l'expression d'attitudes sociales par des Agents Conversationnels Animés.

Le projet Demeanour a exploré l'utilisation de personnages virtuels pour l'improvisation théâtrale [Ballin et al., 2004]. Dans ce projet, plusieurs personnes se regroupent en ligne en étant représentées par des avatars dans le but d'improviser une histoire. Des études passées ont montré que si des utilisateurs doivent eux-même sélectionner le comportement de leur avatar, cela les distrait d'autres tâches et appauvrit globalement l'expérience [Cassell & Thorisson, 1999, Slater et al., 2000]. L'approche de Demeanour est de laisser les utilisateurs définir l'attitude de leur avatar, à partir

du modèle bi-dimensionnel d'Argyle (voir Section 2.2.2), et d'automatiser le comportement des avatars à partir de ces attitudes. En se fondant sur la littérature, les auteurs définissent l'influence de l'attitude sur plusieurs variables : la relaxation (relaxé *vs* nerveux), la quantité d'espace occupé (un personnage dominant prendra plus d'espace), *etc.* Ces paramètres viennent enfin affecter précisément la posture et le regard des avatars. Par exemple, afin d'exprimer une attitude amicale envers son interlocuteur, la posture choisie pour un avatar va être proche de son interlocuteur et orientée vers lui. Ce modèle se limite cependant à quelques modalités, la posture et le regard, et à un contexte où l'interaction ne se fait pas directement avec un humain.

FIGURE 3.4: Exemples de postures générées dans le projet Demeanour [Ballin et al., 2004].

Fukayama *et al.* ont proposé un modèle de regard dans le but de contrôler les impressions renvoyées par des agents, c'est à dire comment ces agents sont évalués par des utilisateurs sur différentes mesures sociales et relationnelles (*ex.* attentionné, attentif, chaleureux, tolérant, *etc.*) [Fukayama et al., 2002]. Le modèle proposé est un modèle de Markov à deux états : un état où le regard est dirigé vers l'utilisateur, et un second état où le regard est détourné. Les paramètres du modèle sont obtenus à partir de travaux existants [Kendon, 1967, Cook & Smith, 1975, Argyle & Cook, 1976, Waxer, 1977] : quantité totale de regards dirigés vers l'utilisateur, durée moyenne des fixations (*i.e.* instants où le regard est fixe), directions de détournement du regard. Les auteurs réalisent une étude où des participants observent des vidéos où uniquement une paire d'yeux est représentée. Différentes conditions sont assignées aux participants, correspondant à différents paramétrages du modèle. Les impressions des participants sont évaluées avec une échelle différentielle sémantique (*i.e.* un ensemble de paires d'adjectifs opposés *ex. amical - inamical*). Les auteurs procèdent à une analyse factorielle afin d'évaluer les facteurs différenciant le plus les impressions récoltées par les participants. Ils analysent ensuite les deux facteurs les plus importants. Le premier est un facteur proche de l'amicalité (corrélation de 0.81

avec l'échelle *unfriendly* - *friendly*, 0.77 avec *cold* - *warm*), et le deuxième facteur est proche de la dominance (corrélation de 0.84 avec l'échelle *unassured* - *assured*, 0.83 avec *weak* - *strong*). Ils montrent que l'agent donne une impression moins amicale lorsque la quantité de regards dirigés vers l'utilisateur est trop faible (*i.e.* 25%) ou trop forte (*i.e.* 100%). L'impression de dominance est corrélée avec la quantité de regards dirigés vers l'interlocuteur. Les détournements de regards vers le bas sont jugés moins dominants, et les regards vers le coté sont jugés moins amicaux. Ces résultats sont limités car seuls les yeux de l'agent étaient affichés, et les résultats obtenus pourraient être différents avec des Agents Conversationnels complets. Enfin, le modèle n'est pas directement utilisable dans un Agent Conversationnel Animé dédié à l'interaction car seule la fonction expressive du regard est considérée, et pas les autres fonctions du regard comme la régulation de l'interaction [Kendon, 1967].

L'agent relationnel Laura [Bickmore & Picard, 2005] a été développé dans le but de nouer des relations à long terme avec des utilisateurs. Laura essayait de motiver des utilisateurs à adopter une activité sportive. Lors d'une étude réalisée sur une durée d'un mois, des utilisateurs ont interagi avec Laura chaque jour afin de discuter de leurs progrès dans la reprise d'une activité sportive. Les dialogues étaient écrits à l'avance, et les utilisateurs pouvaient choisir leurs phrases à l'aide d'un menu. Trois conditions étaient étudiées : un groupe de participants n'interagissait pas avec l'agent (contrôle), un groupe interagissait avec le système paramétré pour être neutre (non-relationnel) et le dernier groupe interagissait avec le système paramétré pour être plus amical (relationnel). Dans le cas où l'agent devait être plus amical, l'agent était paramétré suivant une analyse de travaux existants : il produisait plus de gestes, de mouvements de tête et d'expressions faciales d'émotions (*ex.* empathie lorsque l'utilisateur exprime des difficultés), et se tenait plus près de l'utilisateur sur l'écran. Les auteurs observent que les participants interagissant avec l'agent relationnel lui attribuent de meilleurs scores sur des mesures d'appréciation, de confiance et de respect. Cependant, cette préférence à interagir avec l'agent ne se traduit pas par un plus grand retour à une activité sportive. La limite principale de Laura est que le comportement et les phrases du système sont définies à l'avance (scriptées). De plus, la dominance exprimée par l'agent n'était pas étudiée. Enfin, les utilisateurs n'interagissaient pas de manière naturelle avec Laura (*i.e.* par des conversations en langage naturel), mais par le biais d'un menu.

Bee *et al.* ont étudié la dominance exprimée par un personnage virtuel dans une série d'études [Bee et al., 2009, Bee et al., 2010]. Dans la première étude [Bee et al., 2009], ils ont étudié l'impact de différentes expressions d'émotions, de différentes directions de regard et de différentes positions de tête sur l'expression de la dominance. Ils montrent qu'un agent exprimant des émotions de joie, de colère et de dégoût est évalué plus dominant qu'un agent avec une expression neutre, tandis qu'un agent exprimant la tristesse ou la peur est jugé moins dominant qu'un agent ayant une expression neutre. De plus, si l'agent oriente sa tête vers le haut, il est évalué plus dominant que s'il oriente sa tête vers le bas. Enfin, un évitement de regard diminue la perception de dominance lors d'expressions de joie, l'augmente lors d'émotions de colère et de peur, et n'a pas d'effet lorsqu'il accompagne des expressions de dégoût, de tristesse, ou une expression neutre. Dans leur seconde étude, Bee *et al.* ont étudié l'attitude exprimée par un agent doté d'un modèle de dialogue permettant l'expression de différentes personnalités (*i.e.* introverti ou extraverti, désagréable ou agréable) et d'un modèle de comportement de regard [Bee et al., 2010]. Ils observent que ni la modalité verbale ni la modalité non-verbale ne domine l'autre. En effet, l'attitude de l'agent est bien perçue lorsque les modèles sont utilisés de manière complémentaire (*ex.* le regard est droit et la tête est haute pour exprimer de la dominance, la phrase est choisie pour exprimer de l'extraversion). Mais lorsque les modèles de dialogue et de regard sont utilisés de manière contradictoire, alors la perception de l'agent dépend des traits exprimés (*ex.* la dominance exprimée non-verbalement n'est pas reconnue lorsque c'est une personnalité agréable qui est exprimée par le modèle de dialogue).

De nombreux travaux en modélisation de comportement pour Agents Conversationnels Animés se sont concentrés sur le comportement d'un agent locuteur (*i.e.* un agent prenant la parole) ou interlocuteur (*i.e.* un agent à qui on parle directement). Mais lors de conversations de groupes, un agent peut aussi participer sans être le locuteur ou l'interlocuteur direct du locuteur (*side participants*). Des agents peuvent aussi être présents dans l'environnement virtuel sans participer directement à la conversation (*bystanders*). Dans le projet Gunslinger, Lee et Marsella ont modélisé le comportement d'agents adoptant ces rôles secondaires à partir des dimensions d'attitudes d'Argyle (amicalité, dominance) [Lee & Marsella, 2011]. Afin de collecter des données sur les comportements de tels personnages, des séances d'improvisation ont été réalisées par des volontaires à partir d'un scénario ayant lieu dans un sa-

loon du Far West. Ce scénario met en scène un shériff, un hors-la-loi, une fille de joie et un barman. Les conversations du scénario permettent de mettre en scène les différents rôles d'une conversation, ainsi que différentes attitudes : par exemple, le hors-la-loi (*i.e.* locuteur, hostile envers le shériff) menace la fille de joie (*i.e.* interlocutrice, amicale envers le shériff) en présence du shériff (*i.e. side participant*, dominant envers le hors-la-loi), tandis que le barman se tient en retrait (*i.e. bystander*, amical envers le shériff). A partir des comportements observés des acteurs, les auteurs extraient des règles qui permettent de choisir des comportements appropriés pour des personnages secondaires, en fonction de leurs intentions communicatives, et de leurs attitudes envers les autres personnages de l'interaction. Par exemple, une règle définie indique que lorsqu'un locuteur dominant et hostile donne un ordre à son interlocuteur, alors les personnages secondaires ayant une attitude soumise envers ce locuteur vont adopter une posture recroquevillée et faire un pas en arrière. Ces règles ont été implémentées dans un système permettant de jouer le scénario du saloon avec des Agents Conversationnels Animés (voir Figure 3.5).

FIGURE 3.5: Personnages du scénario Gunslinger.

Un modèle pour l'expression d'attitudes et de personnalités par des Agents Conversationnels Animés lors de salutations a été proposé par Cafaro *et al.* [Cafaro et al., 2012]. Le modèle computationnel proposé est fondé sur les travaux de Kendon sur les différentes phases de salutations [Kendon, 1990] et sur les travaux de Hall sur

la proxémie [Hall, 1966]. A partir de ces travaux, les auteurs définissent des distances où le comportement de l'agent est manipulé afin d'exprimer de l'amicalité et de l'extraversion. Les auteurs choisissent les signaux non-verbaux de l'agent en se fondant sur la littérature concernant l'expression d'attitudes et de personnalités [Argyle, 1988, Burgoon et al., 1984, Richmond & McCroskey, 2000]. Ils considèrent le sourire, la quantité de regard et la distance de l'agent à l'utilisateur. Les auteurs réalisent une expérience pour étudier l'impact de ces trois signaux sur l'attitude et la personnalité exprimée par un agent lors de salutations. Les participants observent des vidéos où la caméra se rapproche d'un agent dont le comportement est affecté par les trois variables suivantes. Tout d'abord, lorsque la caméra s'approche à huit mètres de l'agent (phase de *salutation distante*), l'agent va sourire ou non. Ensuite, à partir de cinq mètres, l'agent va regarder l'utilisateur pendant deux secondes, ou détourner le regard. Enfin, à partir de trois mètres (phase d'*approche*), l'agent va pouvoir faire un pas en direction de l'utilisateur. L'étude montre qu'un agent souriant et regardant plus longuement l'utilisateur est perçu plus amical, et que l'agent se rapprochant est perçu plus extraverti. Ces travaux sont toutefois limités aux situations de salutations et à l'expression de l'amicalité.

Ravenet *et al.* ont utilisé une méthode de *crowdsourcing* (littéralement « approvisionnement de données par la foule ») pour construire un modèle computationnel d'expression d'attitudes sociales [Ravenet et al., 2013]. Une interface web a été développée sur laquelle des utilisateurs pouvaient choisir les signaux non-verbaux produits par un agent (*ex.* gestes de plusieurs amplitudes et intensités, évitement ou non de regard, mouvements de tête, *etc.*, voir Figure 3.6). Chaque participant se connectant sur cette interface devait choisir les signaux non-verbaux de l'agent par rapport à une série de consignes (*ex.* « Quels signaux l'agent doit choisir pour *poser une question* en exprimant une *attitude dominante* ? »). Les auteurs ont ainsi collecté des données sur les signaux non-verbaux qu'un Agent Conversationnel Animé peut produire afin d'exprimer une attitude en même temps qu'une intention communicative. Ils ont ensuite proposé un réseau Bayésien afin de modéliser l'influence de l'attitude exprimée et l'intention communiquée sur les signaux non-verbaux choisis. Ce réseau est paramétré à partir des données récoltées par *crowdsourcing*. Le modèle a ensuite été combiné avec un modèle de dialogue pour l'expression d'attitudes, et évalué pour l'expression de l'amicalité et de l'hostilité [Callejas et al., 2014]. Les résultats de l'étude montrent que le modèle de dialogue seul ne parvient pas à

exprimer les attitudes d'amicalité et d'hostilité. En revanche, le modèle de comportement non-verbal seul parvient à exprimer l'amicalité, mais pas l'hostilité. Enfin, la combinaison des deux modèles parvient à exprimer les deux attitudes. Une des forces de l'approche proposée est le réseau Bayésien, qui permet de modéliser le lien entre les attitudes et les signaux non-verbaux sous forme de probabilités. Cela implique que pour une attitude et intention données en entrée, le modèle peut proposer plusieurs animations variées. Cependant, plusieurs limites existent sur ces travaux. Tout d'abord, seules les attitudes d'amicalité et d'hostilité ont été évaluées, et non la dominance et la soumission. Certaines modalités ne sont pas considérées (*ex.* la posture). De plus, le modèle proposé ne considère qu'une intention communicative à la fois et ne prend pas en compte la séquentialité des signaux non-verbaux. Si plusieurs intentions communicatives sont présentes l'une après l'autre dans la même phrase, alors le choix des signaux non-verbaux pour exprimer une des intentions sera fait indépendamment du choix des signaux non-verbaux choisis pour les autres intentions. L'algorithme retournera alors une séquence de signaux non-verbaux donc les interactions entre signaux n'ont pas été considérées. Enfin, le modèle a été évalué par le biais de vidéos et pas en interaction, et il se peut que les impressions des participants soient différentes dans un contexte d'interaction.

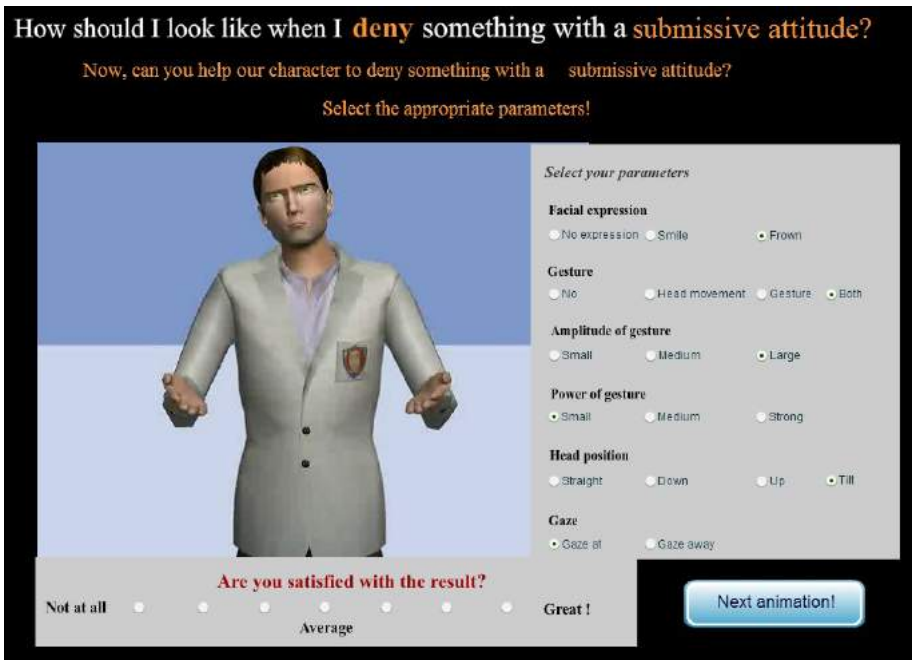


FIGURE 3.6: Interface de collecte de comportements par *crowdsourcing* utilisée dans [Ravenet et al., 2013].

Pour résumer, les modèles que nous avons présentés possèdent certaines limites. La plupart ne considèrent qu'un nombre limité de modalités [Fukayama et al., 2002, Ballin et al., 2004, Bickmore & Picard, 2005, Bee et al., 2009, Bee et al., 2010, Ravenet et al., 2013]. Certains considèrent des rôles secondaires seulement [Lee & Marsella, 2011] ou des situations particulières (*ex.* les salutations [Cafaro, 2014]). Enfin, aucun des modèles existants ne considère des séquences de signaux non-verbaux dans l'expression des attitudes. Dans notre travail de thèse, nous avons étudié la planification de séquences de signaux non-verbaux et pas uniquement de signaux non-verbaux indépendants les uns des autres. En effet, le sens de signaux non-verbaux peut être modifié dans le cadre d'une séquence de signaux non-verbaux (voir Section 2.3). De plus, les attitudes ne s'expriment pas à un instant donné mais affectent le comportement sur de plus longues périodes, et planifier des séquences de signaux non-verbaux permet de s'assurer de la cohérence des différents signaux qui les constituent par rapport à l'attitude à exprimer. Dans cette perspective, nous présentons un état de l'art sur la planification de séquences de signaux non-verbaux dans la section suivante.

3.3 PLANIFICATION DE SÉQUENCES DE SIGNAUX NON-VERBAUX

L'approche la plus répandue dans la planification de comportement d'Agents Conversationnels Animés consiste à définir un lexique contenant, pour chaque intention communicative considérée, la liste des signaux non-verbaux qui peuvent l'exprimer [Cassell et al., 2004, Poggi et al., 2005, Lee & Marsella, 2006, Mancini & Pelachaud, 2007, Hartholt et al., 2013, Cafaro, 2014]. Les intentions communicatives (*ex.* poser une question, mettre l'accent sur un mot) sont traitées indépendamment, et les signaux non-verbaux choisis pour les réaliser sont sélectionnés indépendamment des signaux précédents ou suivants. Or, comme nous l'avons montré dans la section 2.3, le sens d'un signal non-verbal peut être affecté par les signaux proches [Keltner, 1995, Burgoon & Le Poire, 1999, With & Kaiser, 2011, Jack et al., 2014].

Des modèles séquentiels ont été appliqués avec succès pour la synthèse d'animation de différentes modalités d'Agents Conversationnels Animés, par exemple pour la synthèse de mouvements de lèvres synchronisés avec la parole (*lip-sync*) [Tamura et al., 1998, Gregor & Richmond, 2010]. Ding *et al.* ont proposé d'utiliser des modèles de Markov cachés pour la création d'animations de mouvements de sourcils

et de tête à partir de fichiers audio de personnes parlant sous le coup de l'émotion [Ding et al., 2013], et pour la création d'animations de rire comprenant mouvements des lèvres, de la mâchoire, de la tête, des sourcils, du torse et des épaules, à partir de fichiers audio de rire [Ding et al., 2014]. Dans ce cadre, chaque élément de la séquence consiste en un ensemble de paramètres d'animations pour les modalités concernées, et l'utilisation d'un modèle séquentiel dont les paramètres sont appris automatiquement permet d'obtenir des animations lisses et sans discontinuités. Ces approches présentent le désavantage de fonctionner au niveau des paramètres d'animation, ce qui rend leurs résultats difficiles à interpréter en termes de signaux non-verbaux et d'intentions communicatives.

Dans cette section, nous présentons des modèles de planification de signaux non-verbaux pour Agents Conversationnels Animés qui utilisent des représentations séquentielles. Nous introduisons d'abord des modèles de planification de signaux non-verbaux conversationnels (*ex.* gestes co-verbaux, hochements de tête d'un locuteur). Ensuite, nous évoquons des approches qui ont modélisé l'expression d'émotions par des séquences de signaux non-verbaux. Bien que nous ne travaillons pas sur l'expression d'émotions dans notre thèse, les représentations des signaux non-verbaux et de leurs relations que ces approches ont utilisées sont pertinentes à mentionner car celles-ci pourraient être réutilisées dans notre cadre.

3.3.1 Signaux non-verbaux conversationnels

Lee et Marsella ont proposé un modèle pour prédire des moments adéquats pour la production de hochements de tête lors de discours à partir de modèles de Markov cachés [Lee & Marsella, 2010]. Les entrées de ces modèles de Markov cachés sont la séquence de mots que l'agent va prononcer, accompagnés de caractéristiques linguistiques (*ex.* partie de discours, étiquette émotionnelle, début ou fin de phrase, *etc.*). A partir d'un corpus annoté, ils ont entrainé la prédiction de hochements de tête sur des trigrammes, *i.e.* des séquences de trois mots. Si la représentation par modèles de Markov cachés permet effectivement de raisonner sur des séquences, seule la relation séquentielle entre des mots successifs est modélisée, et pas les relations entre les signaux non-verbaux successifs. De plus, aucun autre signal que le hochement de tête n'est considéré.

Marsella *et al.* ont présenté Cerebella [Marsella et al., 2013], système permettant de détecter automatiquement les fonctions communicatives contenues dans un texte et un fichier audio correspondant. Plusieurs types d'analyses sont réalisées sur le texte (rhétorique, syntaxique, lexicale, sémantique) et sur le fichier audio (détection du niveau d'agitation) afin d'identifier les fonctions communicatives du locuteur. Ces analyses utilisent des règles définies à partir de la littérature sur les fonctions de la communication. Après ces analyses, les fonctions communicatives sont transformées en signaux non-verbaux, encore une fois à l'aide de règles définies manuellement. Certaines de ces règles considèrent la relation séquentielle entre plusieurs signaux non-verbaux. Par exemple, pour exprimer un *contraste* entre deux éléments du discours, il est possible de réaliser deux gestes similaires, l'un avec le bras gauche, l'autre avec le bras droit. La limite principale de Cerebella est ainsi que ce système repose essentiellement sur un ensemble de règles définies manuellement.

Xu *et al.* ont proposé un modèle pour la planification de gestes [Xu et al., 2014], fondé sur la théorie de Calbris [Calbris, 2011]. Dans cette théorie, si une relation existe entre plusieurs idées atomiques exprimées par le discours et le comportement non-verbal, alors cet ensemble d'idées constitue une *unité idéationnelle*. Par exemple, « (*Je veux bien parler de tout*) sauf de (*ce que veut mon mari*) » : ici, les deux idées entre parenthèses sont liées et constituent une unité idéationnelle. Les gestes utilisés pour exprimer différents concepts d'une même unité idéationnelle possèdent des caractéristiques communes de forme et de rythme. Ces similitudes explicitent la relation entre les différents concepts exprimés par plusieurs gestes se succédant. Inversement, un changement de sujet, c'est à dire un changement d'unité idéationnelle, est illustré par une rupture dans la cohérence de forme et de rythme de gestes successifs. Xu *et al.* définissent un algorithme de planification de gestes à partir de phrases définies dans le format FML [Heylen et al., 2008]. Ces phrases sont enrichies de balises indiquant les limites des différentes unités idéationnelles qu'elles contiennent. L'algorithme proposé fonctionne à partir d'un ensemble de contraintes opérant sur les séquences de gestes, selon que ces gestes appartiennent à la même unité idéationnelle ou non (*ex.* les mains se relaxent à la fin de gestes à la fin d'une unité idéationnelle, mais ne le font pas entre des gestes de la même unité idéationnelle). Ces travaux et la théorie sur laquelle ils s'appuient sont cependant limités à la modalité gestuelle.

Dans la prochaine section, nous présentons des approches ayant modélisé l'expression

d'émotions par des séquences de signaux non-verbaux.

3.3.2 Expressions d'émotions

Pan *et al.* ont proposé une approche pour la modélisation de séquences de comportements utilisant des graphes de mouvement (*motion graphs*) [Pan et al., 2007]. A partir d'un corpus d'animations (*ex.* obtenues par capture de mouvement), cette méthode consiste à automatiquement construire un graphe orienté, où les arcs correspondent à des segments d'animations, et où les sommets correspondent à des instants où une transition peut être réalisée entre deux segments d'animations. Parcourir ce graphe permet de créer une nouvelle animation à partir de segments provenant d'animations différentes. Pan *et al.* ont appliqué cette méthode à la génération d'animation de mouvements de tête et d'expressions faciales pour l'expression d'états mentaux complexes (*ex.* intérêt, concentration, désaccord, *etc.*). Leur approche a consisté à récolter un corpus de vidéos où des acteurs expriment les états mentaux considérés. Une méthodologie d'analyse automatique de vidéos [el Kaliouby & Robinson, 2005] a ensuite été utilisée pour extraire automatiquement des paramètres décrivant l'expression faciale et les mouvements de la tête de l'acteur à chaque image (*ex.* hochement de tête, haussement de lèvre, haussement de sourcils, *etc.*). Ces paramètres peuvent être utilisés directement pour animer la tête et le visage afin de reproduire directement l'expression de l'acteur dans la vidéo d'origine. A partir de ces animations, les auteurs construisent un graphe de mouvement pour chaque état mental des vidéos de leur corpus. En suivant différents parcours dans ces graphes, plusieurs animations variées peuvent être créées pour un même état mental. Ces travaux sont cependant limités par la méthode d'extraction automatique de paramètres des vidéos, qui ne considère pas tous les mouvements possibles du visage (*ex.* l'ouverture de la bouche n'est pas considérée). De plus, les animations générées n'ont pas été évaluées.

Paleari *et al.* [Paleari et al., 2007], Malatesta *et al.* [Malatesta et al., 2007] et Courgeon *et al.* [Courgeon et al., 2014] ont défini des expressions faciales séquentielles d'émotions à partir des prédictions de la théorie de l'évaluation cognitive de Scherer [Scherer & Ellgring, 2007]. Dans le modèle de Scherer [Scherer & Ellgring, 2007], plusieurs évaluations cognitives successives sont réalisées, sur la *pertinence* d'un évènement (*ex.* est-il nouveau, soudain ? Est-il intrinsèquement agréable ?), sur son

FIGURE 3.7: En haut, vidéo originale montrant l'émotion d'incertitude. En bas, animation générée par le modèle de Pan *et al.* pour l'expression d'incertitude [Pan et al., 2007].

degré d'implication (*ex.* à qui est-ce la faute ? Est-il lié à mes buts ?), *etc.* Ces évaluations ont lieu les unes après les autres et chacune influence les expressions faciales. Une expression faciale d'émotion devient ainsi naturellement un processus séquentiel, et ce sont les dynamiques de l'expression qui permettent d'identifier l'émotion ressentie. Palerai *et al.* ont considéré les émotions de colère, de dégoût, de peur, de joie et de tristesse [Paleari et al., 2007]. Ils ont défini manuellement l'intensité et les contraintes temporelles entre tous les signaux faciaux de ces émotions. Malatesta *et al.* réalisent quand à eux des animations additives : chaque évaluation cognitive active de nouvelles *Action Units* (combinaisons de mouvements musculaires faciaux, unités de mesure du schéma de codage d'expressions faciales FACS [Ekman & Friesen, 1977]), qui viennent s'ajouter à l'expression faciale affichée jusque là [Malatesta et al., 2007]. L'Agent Conversationnel Animé MARC peut prendre le rôle d'un adversaire au jeu de plateau Reversi [Courgeon et al., 2014]. Celui-ci est doté d'un module d'expression d'émotions suivant plusieurs modèles, dont le modèle CPM d'évaluation cognitive proposé par Scherer [Scherer & Ellgring, 2007]. Dans ce dernier cas, plusieurs évaluations cognitives successives en réaction au jeu de l'utilisateur (*ex.* le coup de l'utilisateur est-il inattendu ? Est-il favorable ou défavorable à MARC ?) sont exprimées par une séquence d'expressions faciales (voir Figure 3.8).

Niewiadomski *et al.* ont proposé une représentation pour des expressions d'émotions par des séquences de signaux multimodaux [Niewiadomski et al., 2011]. Cette représentation est fondée sur un ensemble de contraintes temporelles entre les signaux,

FIGURE 3.8: Expression de la tristesse par MARC. En haut, un modèle suivant la théorie des émotions discrètes est utilisé [Ekman, 1972]. En bas, des évaluations cognitives suivant la théorie de Scherer mènent à une expression de tristesse [Courgeon et al., 2014].

telles que $Signal_1$ *précède* $Signal_2$ ou $Signal_1$ *et* $Signal_2$ *ont lieu simultanément*, et de contraintes sur l'absence ou la présence de signaux, telles que $Signal_1$ *ne peut être présent sans* $Signal_2$. Les auteurs ont annoté un ensemble de vidéos comprenant des expressions d'émotions afin d'analyser les signaux non-verbaux apparaissant lors de l'expression de ces émotions. Les signaux annotés sont le regard, les expressions faciales (codées avec le schéma de codage *FACS* [Ekman & Friesen, 1977]), les mouvements de tête, les gestes et les mouvements de torse. A partir de ces annotations, ils définissent manuellement les relations temporelles entre les signaux non-verbaux exprimant les émotions de leur corpus. Un algorithme est proposé afin de pouvoir générer de nouvelles animations de différentes durées à partir d'une seule représentation de la même émotion. Une évaluation est ensuite réalisée afin de comparer les séquences multimodales générées avec leur algorithme à des images fixes et à des séquences multimodales où l'ordre des signaux est modifié. Les résultats montrent que, pour les huit émotions considérées (Colère, anxiété, gaieté, embarras, peur panique, fierté, tension, soulagement), les séquences multimodales générées avec l'algorithme proposé présentent les meilleurs taux de reconnaissance dans sept des cas (seule l'émotion de gaieté ou *Cheerfulness* est mieux reconnue dans le cas d'une image fixe montrant le sourire). Ces résultats confirment que l'ordre des signaux d'une séquence

influe sur l'interprétation de cette séquence de signaux non-verbaux. La limitation majeure de ce travail est que les contraintes temporelles sont définies manuellement et pas par apprentissage automatique.

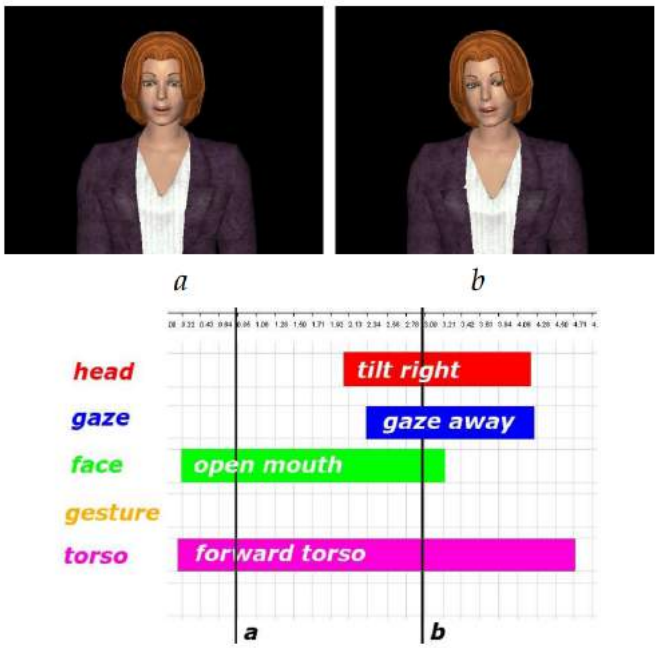


FIGURE 3.9: Expression multimodale de soulagement obtenue avec le modèle de [Niewiadomski et al., 2011].

Les approches que nous avons présentées qui considèrent des séquences de signaux non-verbaux pour la modélisation du comportement d'ACAs sont cependant limitées. De nombreux modèles ont été proposés pour la synthèse d'animations qui utilisent des modèles de Markov cachés pour tirer partie de leur représentation séquentielle. Cependant, ces approches se placent au niveau des paramètres d'animation et pas des signaux non-verbaux. Les approches de Lee et Marsella [Lee & Marsella, 2010] et de Pan *et al.* [Pan et al., 2007] ont l'avantage d'être automatiques, mais elles ne concernent qu'un ensemble de signaux limité. Les travaux de Niewiadomski *et al.* [Niewiadomski et al., 2011] sont particulièrement intéressants car la planification se place au niveau du signal non-verbal, et raisonne sur les relations temporelles entre les signaux. Cependant, cette méthode repose sur un processus d'analyse et de représentation manuel.

CONCLUSION

Dans ce chapitre, nous avons présenté un état de l'art sur les Agents Conversationnels Animés selon trois thématiques. Tout d'abord, nous avons donné un aperçu des applications pour l'entrainement social dans lequel des Agents Conversationnels Animés sont utilisés. Nous nous sommes particulièrement intéressés aux systèmes d'audiences virtuelles, qui ont jusqu'ici été utilisées pour la réduction de l'anxiété liée aux situations de prise de parole en public, et aux systèmes simulant des entretiens d'embauche. La limite majeure partagée par les systèmes que nous avons présentés est que ceux-ci ne sont pas interactifs, et qu'ils ne modélisent pas les aspects socio-émotionnels des situations sociales qu'ils simulent. En particulier, dans le cadre de la simulation d'entretiens d'embauches, les recruteurs devraient pouvoir exprimer différentes attitudes.

Nous avons aussi détaillé les travaux ayant étudié l'expression d'attitudes sociales par des Agents Conversationnels Animés. Ces travaux considèrent des ensembles de modalités et de signaux non-verbaux variables, mais jamais l'ensemble des modalités non-verbales. Certains sont aussi limités à des situations particulières : par exemple, l'expression d'amicalité lors de salutations [Cafaro, 2014], ou la modélisation du comportement des personnages secondaires présents dans des environnements virtuels [Lee & Marsella, 2011]. Enfin, aucun de ces travaux n'a considéré des séquences de signaux non-verbaux.

La dernière thématique que nous avons abordée est la planification de séquences de signaux non-verbaux. Si des modèles séquentiels ont été utilisés avec succès dans le domaine de la synthèse d'animations (*ex.* modèles de Markov cachés), les approches planifiant des séquences de signaux non-verbaux sont encore rares. Nous avons identifié deux domaines dans lequel la planification de séquences de signaux non-verbaux a particulièrement été étudiée : les séquences de signaux non-verbaux conversationnels, comme les séquences de gestes liés, et les séquences de signaux non-verbaux pour l'expression d'émotions. Ces approches sont limitées à un nombre de modalités restreintes, à l'exception des travaux de Niewiadomski *et al.* [Niewiadomski et al., 2011]. Enfin, la plupart de ces approches reposent sur un travail manuel de définition de règles ou de contraintes.

Ces constats ont guidé l'orientation de notre travail de thèse. Notre objectif principal

est de proposer un modèle pour la planification de séquences de signaux non-verbaux pour l'expression d'attitudes sociales. La littérature sur l'expression des attitudes sociales (Chapitre 2) ne nous fournit pas d'informations sur les attitudes exprimées par des séquences de signaux non-verbaux, et nous avons donc opté pour une approche d'extraction automatique de connaissances (Chapitre 5) à partir d'un corpus multimodal (Chapitre 4). Nous avons proposé un modèle de planification de comportement (Chapitre 6), que nous avons ensuite utilisé pour implémenter un recruteur virtuel adaptant ses attitudes par rapport à la performance de l'utilisateur (Chapitre 7). Enfin, nous avons aussi proposé et évalué un système d'audience virtuelle interactive fournissant un retour immédiat à l'utilisateur sur sa performance par le biais du comportement non-verbal de l'audience (Chapitre 8).

Synthèse du chapitre

1. Des systèmes d'entrainement virtuel pour l'amélioration des compétences interpersonnelles ont été proposés dans de nombreux domaines. Cependant, les applications se concentrant sur l'amélioration du comportement non-verbal sont encore rares, et ces applications ne produisent pas de retour en temps réel à l'utilisateur sur sa performance.
2. Des modèles d'expression d'attitudes sociales pour des Agents Conversationnels Animés ont été proposés. Ces modèles ne considèrent pas toutes les modalités du comportement non-verbal, ou bien sont utilisés dans un contexte particulier (*ex.* salutations, personnages secondaires). Enfin, aucun ne considère l'impact de séquences de signaux non-verbaux dans l'expression de l'attitude.
3. Le sens de signaux non-verbaux peut être affecté par les signaux multimodaux proches. Cependant, les modèles de comportement pour Agents Conversationnels Animés considérant des séquences de signaux non-verbaux et pas uniquement des signaux non-verbaux indépendants sont rares.

4 Corpus multimodal pour la modélisation d'expressions d'attitudes de recruteurs virtuels

Les travaux ayant étudié l'expression de l'attitude par le comportement non-verbal ne se sont pas intéressés aux séquences de signaux non-verbaux. Afin de pouvoir étudier et modéliser l'influence de séquences de signaux non-verbaux sur l'expression d'attitudes, nous avons donc utilisé un corpus multimodal. Une définition générale des corpus multimodaux est proposée par Allwood : « *A multimodal digitized corpus is a computer-based collection of language and communication-related material drawing on more than one sensory modality or on more than one production modality* » [Allwood, 2008]. Cette définition d'un corpus multimodal, qui englobe ainsi toute collection de matériaux de nature audio-visuelles à propos d'un phénomène lié à la communication, est précisée plus loin : « *In a more narrow sense, we might require that the audiovisual material should be accompanied by transcriptions and annotations or codings based on the material* ». De manière plus pratique, Foster et Oberlander définissent un corpus multimodal ainsi : « *A multimodal corpus is an annotated collection of coordinated content on communication channels such as speech, gaze, hand gesture, and body language, and is generally based on recorded human behaviour* » [Foster & Oberlander, 2007]. En résumé, un corpus multimodal est un ensemble de données décrivant le comportement de personnes dans des contenus audio-visuels par un ensemble d'annotations, transcriptions ou codes synchronisés, représentant les signaux d'une ou plusieurs modalités produits par ces personnes (*ex.* leurs gestes, ou leurs expressions faciales).

Les corpus multimodaux sont régulièrement utilisés dans le domaine des Agents Conversationnels Animés. La méthodologie mise en avant par Justine Cassell pour modéliser des ACAs consiste en des itérations successives du cycle *Etude*, *Modélisation*, *Implémentation* et *Test* (voir Figure 4.1) [Cassell, 2007]. On commence par collecter des données (*i.e.* un corpus), qui servent ensuite élaborer un modèle, qui est ensuite implémenté dans un système, qui est enfin testé et évalué afin de cerner

ses limites, et un nouveau cycle peut alors être entamé pour améliorer ce modèle. Rehm et André distinguent deux types d'utilisations principaux des corpus multimodaux [Rehm & André, 2008] : ils peuvent être utilisés directement, par exemple en utilisant des annotations de mouvements faciaux pour générer des expressions faciales pour ACAs, ou alors pour extraire des règles ou paramètres servant ensuite à alimenter des modèles computationnels.

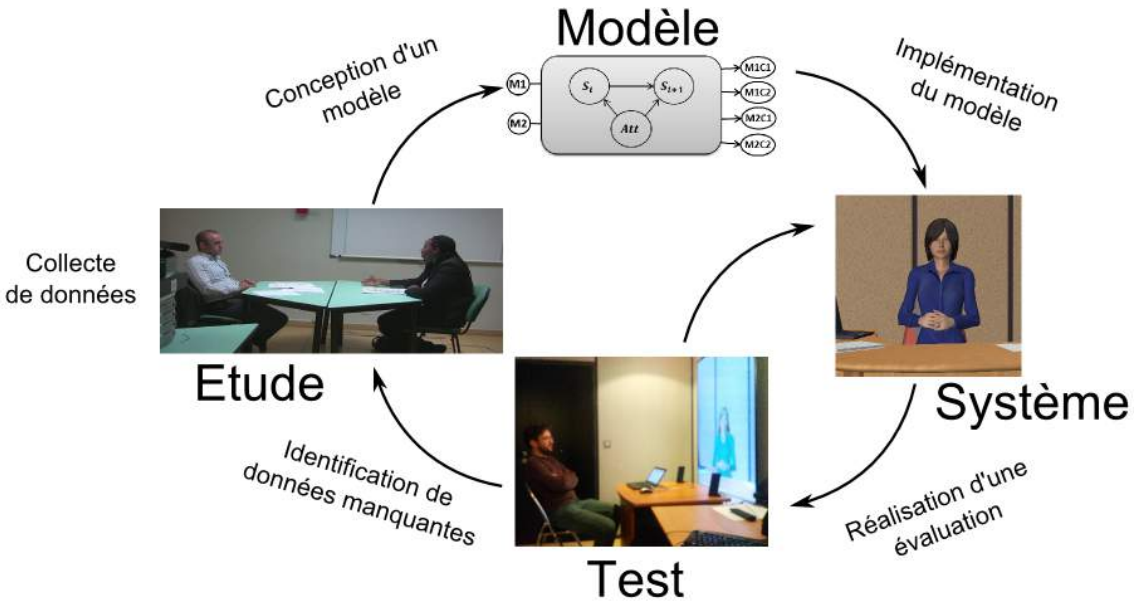


FIGURE 4.1: Méthodologie de modélisation d'Agents Conversationnels Animés proposée dans [Cassell, 2007].

Notre travail suit la seconde approche. Dans ce chapitre, nous présentons le corpus multimodal que nous avons collecté afin d'extraire automatiquement des séquences de signaux non-verbaux exprimant des attitudes. Dans une première section, nous commençons par dresser un aperçu des méthodes et des outils de création de corpus multimodaux et des types d'utilisation de ceux-ci dans le domaine des Agents Conversationnels Animés. Ensuite, nous présentons une sélection de corpus multimodaux existants, et déterminons si ceux-ci peuvent être utilisés pour notre problématique. La troisième section de ce chapitre est dédiée au corpus multimodal que nous avons annoté et utilisé : nous présentons d'abord les caractéristiques générales du corpus et le contexte de leur collecte. Par la suite, nous détaillons le processus d'annotation de ces vidéos réalisé dans cette thèse, annotation du comportement non-verbal d'une part, et de l'attitude d'autre part.

4.1 CRÉATION ET ANNOTATION DE CORPUS MULTIMODAUX

Dans cette section, nous donnons un aperçu des méthodes et des outils de création de corpus multimodaux pour la conception d'Agents Conversationnels Animés.

4.1.1 Création de corpus multimodaux

Lors de la création d'un corpus multimodal ou du choix d'un corpus existant pour l'étude d'un phénomène (*ex.* une émotion ou une attitude), il est important de garder à l'esprit certaines considérations [Picard, 1997, Gunes & Piccardi, 2006] :

- Données spontanées ou posées : les sujets enregistrés dans le corpus expriment-ils le phénomène étudié sur demande (*ex.* acteurs), par le biais d'une méthode d'induction [Cowie et al., 2011], ou bien le phénomène étudié survient-il spontanément ?
- Laboratoire ou conditions réelles : l'enregistrement est-il réalisé dans un environnement contrôlé (dans un laboratoire, avec des conditions d'éclairage contrôlées, sans perturbations ou bruits extérieurs) ou bien dans l'environnement naturel, habituel et non contrôlé des sujets (chez eux ou au travail, sans contraintes sur l'éclairage ou d'autres conditions) ?
- Enregistrement caché ou non : les sujets sont-ils informés qu'ils sont observés et enregistrés ?
- But caché ou non : les sujets savent-ils qu'ils doivent exprimer le phénomène étudié (*ex.* les émotions ou attitudes étudiées) ?

Ces différents aspects peuvent en effet influencer les données obtenues. Par exemple, si des acteurs peuvent exprimer des émotions qui seront bien reconnues par des annotateurs, ces expressions actées ne reflètent pas nécessairement des expressions naturelles d'émotions [Douglas-Cowie et al., 2003]. Cependant, il n'est pas forcément possible de standardiser les conditions d'acquisition de données naturelles, et les données peuvent être bruitées [Douglas-Cowie et al., 2007]. Enfin, les participants conscients d'être observés modifient leur comportement [Labov, 1978, Knight, 2011].

Définir les moyens d'enregistrement (*ex.* caméras, Microsoft Kinect, microphones, *etc.*) et leur quantité est aussi une question décisive, et en lien avec les considérations précédentes [Allwood, 2008]. En effet, si l'objectif du corpus multimodal est d'être utilisé pour pouvoir entraîner un système de reconnaissance automatique d'émotions sur des expressions faciales, alors un bon positionnement des caméras sera crucial (le visage devant être suffisamment bien cadré) et il ne sera probablement pas possible de les cacher aux participants. Inversement, si les données brutes (*ex.* fichiers vidéos, fichiers audio) ne seront pas directement utilisées, mais annotées ou transcrites par la suite par des humains, alors des prises de vues plus distantes seront acceptables.

Enfin, si les données sont transcrites ou interprétées par des experts, par exemple afin de segmenter les instants où un participant semble ressentir une certaine émotion, il faut alors définir un schéma de codage. Un schéma de codage définit des étiquettes permettant de coder les occurrences de phénomènes étudiés sur une ou plusieurs modalités [Dybkjær & Bernsen, 2004]. De nombreux schémas de codage existent pour l'annotation de signaux non-verbaux. Un des premiers est le Facial Action Coding System (FACS), schéma de codage introduit par Ekman et Friesen pour l'annotation des expressions faciales [Ekman & Friesen, 1977]. Le schéma de codage MUMIN (MUltiModal INterfaces [Allwood et al., 2007a]) introduit par Allwood *et al.* se concentre sur la forme générale des comportements multimodaux et de leur fonction. Celui-ci a été conçu pour l'analyse des expressions multimodales dans les phénomènes de *feedback*, de l'étude du tour de parole et de l'organisation de dialogues en séquences de sous-dialogues indépendants [Allwood et al., 2007b]. On peut distinguer deux parties distinctes dans celui-ci, une pour l'annotation des fonctions communicatives, et une pour l'annotation des signaux non-verbaux.

En résumé, la création d'un corpus multimodal nécessite de prêter attention aux conditions de collecte des données (*ex.* données actées ou posées, étude en laboratoire ou « *in the wild* ») et de définir un schéma de codage adapté à l'objectif poursuivi. Une fois ces questions réglées, il faut ensuite décider avec quels outils l'annotation va être réalisée.

4.1.2 Outils d'annotation de corpus multimodaux

De nombreux outils pour l'annotation multimodale sont apparus dans la dernière décennie. Les objectifs, les contextes de création et la qualité de ces outils varient et il appartient au chercheur de choisir l'outil le plus adapté à sa tâche.

On peut distinguer les outils d'annotation selon les types d'évènements annotés : un évènement peut être représenté par une paire valeur-temps (*point-based annotation*). Par exemple, un *changement de posture* a lieu à un instant t. Dans d'autres cas on peut considèrer qu'un évènement possède un début et une fin (*interval-based annotation*), mais une valeur unique. Par exemple, un *sourire* a lieu entre les instants t_1 et t_2. Enfin, d'autres outils permettent d'annoter une ou deux dimensions continues simultanément, de manière continue dans le temps, permettant d'obtenir une courbe de points représentant la valeur d'un phénomène à chaque instant (*trace-based annotation*). On pourra par exemple tracer une courbe décrivant la *valence* de l'émotion ressentie par une personne sur une vidéo.

Un type d'interface très utilisé dans le domaine de l'annotation multimodale est celui des outils dits *track-based* ou *tier-based*. On peut présenter ce type d'outil avec une métaphore de partition musicale, où les différents instruments sont représentés parallèlement (lignes horizontales) et synchronisés temporellement (deux notes alignées verticalement sont simultanées) [Poggi, 2003]. L'idée générale de ce paradigme d'annotation est que les évènements sont annotés à différents niveaux, c'est à dire sur différentes lignes horizontales (*tiers* ou *tracks*) parallèles : la dimension horizontale correspond au temps, et les évènements de nature différente sont annotés sur des niveaux différents, par exemple on peut définir une ligne correspondant aux annotations de gestes, une autre correspondant aux sourires, ou encore à la transcription de la parole.

De nombreux outils d'annotation présentent une interface de ce type, comme EXMARALDA [Schmidt et al., 2011] ou TASX [Milde & Gut, 2002], mais les deux outils les plus répandus sont Elan (voir Figure 4.2) [Wittenburg et al., 2006] et Anvil [Kipp, 2012]. Si ces outils ont des origines différentes, leurs trajectoires de développement les ont amenés à être relativement similaires, et des efforts récents pour rendre ces différents outils interopérables font qu'il est difficile de les départager. Les deux outils supportent la spécification interne ou externe d'un schéma

de codage. Il est possible dans les deux cas de définir des hiérarchies de lignes (*ex.* une ligne « amplitude gestuelle » subordonnée à la ligne « gestes »), et les deux outils proposent différents modes d'entrée de données intuitifs et adaptés à différentes tâches (*ex.* première segmentation rapide mais imprécise, spécification des évènements correspondant aux segments, segmentation précise).

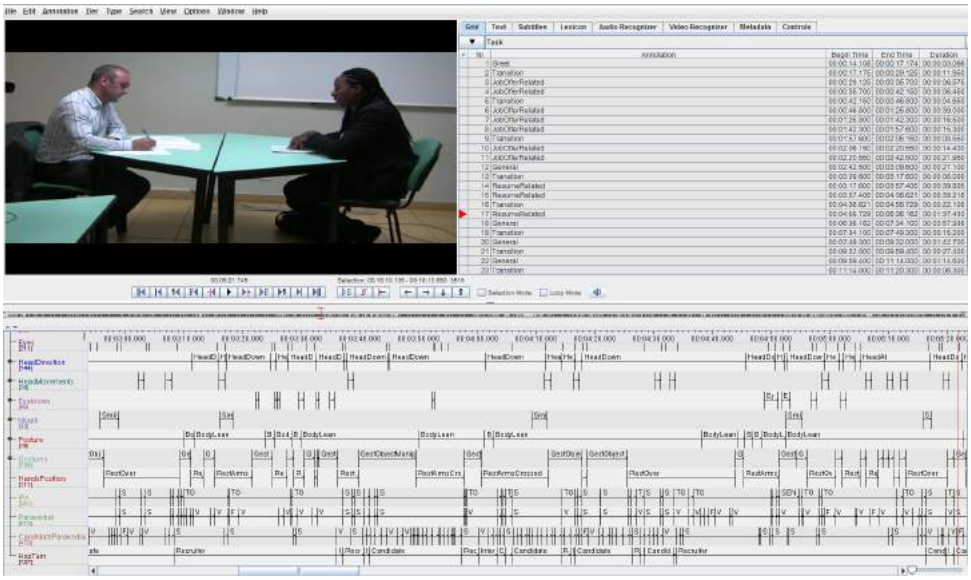


FIGURE 4.2: Environnement d'annotation Elan.

Le paradigme d'annotation continue est plus récent. MoodSwings [Kim et al., 2008] et EmuJoy [Nagel et al., 2007] sont deux outils qui ont été développés pour l'évaluation des dimensions émotionelles de morceaux de musique. Dans le cadre d'enregistrements audio-visuels, Feeltrace [Cowie et al., 2000] et son successeur Gtrace [Cowie et al., 2012] ont été utilisés pour l'annotation d'émotions représentées dimensionnellement dans le cadre des projets HUMAINE[1] et SEMAINE[2]. Dans GTrace, l'interface utilisateur est constituée de deux parties (voir Figure 4.3). A gauche, la vidéo à annoter est affichée. A droite, l'échelle continue d'annotation est présentée. L'utilisateur a le contrôle sur un curseur pouvant être déplacé sur l'échelle d'annotation. La position du curseur sur l'échelle représente la valeur que l'utilisateur attribue au phénomène annoté à l'instant correspondant de la vidéo.

Les annotations continues récoltées avec un outil d'annotation de type *trace* présentent à la fois des avantages et des inconvénients. Comme le présentent Cowie *et*

[1] www.emotion-research.net/projects/humaine
[2] www.semaine-project.eu

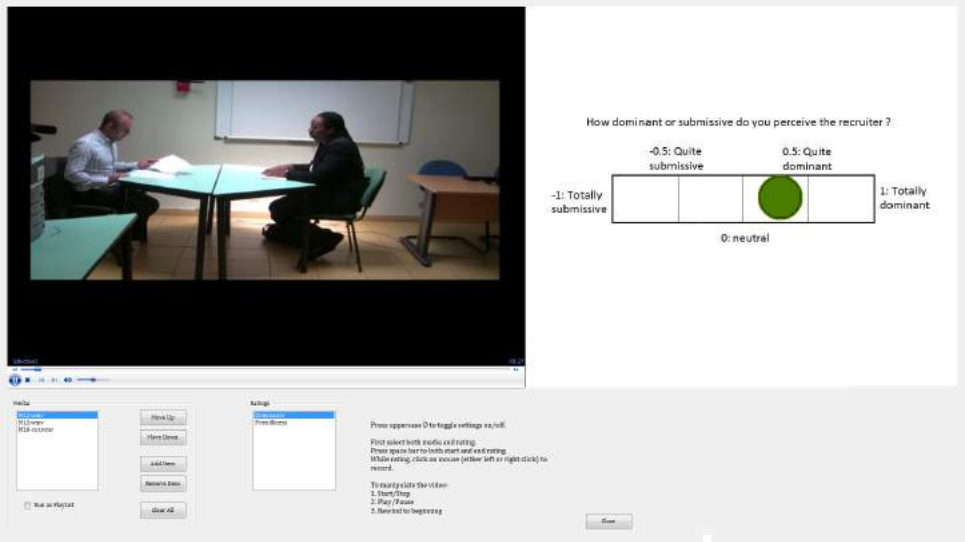


FIGURE 4.3: Environnement d'annotation GTrace.

al., « *The ability to record in real-time has enormous advantages, not least because responses in interactions are likely to be based on instantaneous impressions of an ongoing display [...]. The trade-off is that experimenters have less control* » [Cowie et al., 2012]. Pour résumer, les traces collectées par annotation continue présentent l'avantage de refléter les réactions instantanées des annotateurs et également de permettre la capture des dynamiques des dimensions sous-jacentes à des évènements (*ex.* on peut capturer les variations d'intensité d'émotions dans le temps, plutôt que d'annoter uniquement une étiquette d'émotion sur un intervalle de temps). Les inconvénients principaux de l'annotation continue tiennent aux problèmes de validité inter-annotateurs [Cowie & McKeown, 2010], qui proviennent de sous-problèmes d'échelle, de bruitage des données, et de temps de réaction [Mariooryad & Busso, 2013]. En effet, les annotateurs doivent être attentifs à la vidéo annotée mais aussi simultanément à la position du curseur. La position du curseur peut être imprécise, à la fois parce que les annotateurs peuvent déplacer le curseur en regardant la vidéo et non pas en regardant l'échelle d'annotation, et parce que les natures continues de l'échelle d'annotation comme du phénomène étudié rendent difficile le choix d'une position exacte du curseur (*i.e.* si une des personnes de la vidéo semble très amicale, quelle est la valeur correspondante de cette amicalité sur l'échelle d'annotation ?). Une imprécision temporelle est aussi inévitable car les annotateurs ne réagissent pas instantanément aux stimuli présents sur la vidéo.

4.2 Corpus multimodaux existants

De nombreux corpus multimodaux ont été créés à ce jour dans le cadre de projets de recherche aux objectifs variés. Dans cette section, nous présentons une sélection de corpus particulièrement pertinents par rapport à nos travaux. Pour un état de l'art complet des corpus multimodaux existants, le lecteur peut se référer à [Gunes & Piccardi, 2006, Knight, 2011].

Un des corpus multimodaux parmi les plus larges est le corpus *AMI* (*Augmented Multi party Interaction corpus*) [Carletta et al., 2006], contenant une centaine d'heures de vidéo. Le but du projet AMI était de développer des technologies de support aux réunions. Dans les vidéos du corpus, des groupes de trois à quatre personnes jouent les rôles d'employés d'une équipe travaillant sur un projet de design. Les personnes occupent différentes rôles (*ex.* chef de projet, ou designer). Des caméras, des microphones et plusieurs autres outils spécifiques (tablettes graphiques pour l'écriture, système d'enregistrement des diapositives) enregistrent les interactions de la scène.

Le corpus *Semaine* [McKeown et al., 2010] est une collection de conversations informelles, où un participant interagit avec un opérateur adoptant quatre différentes personnalités. Les interactions sont télé-opérées (les participants étaient dans des pièces différentes), et les vidéos retransmises sont cadrées pour contenir le haut du buste et le visage. Des annotations de dimensions émotionnelles (valence, excitation, pouvoir social) ont été réalisées sur ces interactions.

Le *Corpus of Interactional Data* (CID) [Blache et al., 2010] contient huit dialogues d'une heure où deux personnes discutent de manière informelle. Les participants devaient d'abord évoquer des conflits professionnels, puis des situations insolites. Les participants sont tous les deux équipés d'un micro, et une caméra enregistre l'interaction.

Dans le cadre du projet ANR Comparse, un corpus multimodal a été collecté afin d'étudier le lien entre émotions, stress, et comportement multimodal lors de situations de prise de parole [Giraud et al., 2013]. Les auteurs ont à cette fin adapté le protocole expérimental du *Trier Social Stress Test*, test utilisé dans de nombreuses études afin de déclencher une réaction de stress [Kirschbaum et al., 1993]. Après une première phase de contrôle où les participants lisaient un texte à haute voix,

ceux-ci devaient se présenter en cinq minutes devant un jury de deux personnes, de la même manière que s'ils passaient un entretien d'embauche. Ensuite, une séance de questions et réponses de cinq minutes également était menée par le jury. Le comportement du jury était enregistré à l'aide d'une caméra, tandis que de nombreux capteurs étaient utilisés pour enregistrer le comportement et l'état affectif des participants (Caméra, microphone, Kinect, plate-forme de force, capteurs physiologiques). De plus, des échantillons salivaires étaient collectés à des instants précis de l'expérience. Ceux-ci ont été utilisés pour analyser les concentrations d'hormones liées au stress [Hua et al., 2014]. Ce corpus n'a cependant été disponible qu'à partir de 2013 et n'a donc pas pu être exploité dans le cadre du présent travail de thèse.

Un corpus extrêmement intéressant pour nos travaux est le corpus *HuComTech* [Szekrényes, István, 2014], qui consiste en une cinquantaine d'heures d'interactions dyadiques entre personnes de nationalité hongroise. Les interactions sont divisées en trois parties. Dans la troisième partie, l'un des deux participants pose des questions typiques d'entretiens d'embauche à son interlocuteur. Malheureusement, ce corpus n'a été disponible lui aussi qu'à partir de 2013. Au demeurant, si les interactions étaient bien guidées sous la forme d'entretiens d'embauches, les participants n'avaient pas la consigne d'adopter des attitudes particulières, et n'étant pas des recruteurs professionnels, on peut se demander si leur comportement est bien représentatif des comportements adoptés par de véritables recruteurs.

Après avoir passé en revue différents corpus pertinents existants, nous les avons évalués selon les critères suivants :

- **Attitudes :** les situations d'interaction enregistrées sont elles propices à l'expression d'attitudes interpersonnelles ?
- **Contexte :** les situations d'interaction enregistrées sont elles des entretiens d'embauche (joués ou non), ou au moins des discussions formelles ?
- **Modalités :** les caméras enregistrant les interactions permettent-elles d'observer toutes les modalités impliquées dans l'expression d'attitudes (*ex.* suffisamment éloignées pour observer la posture) ?
- **Dyades :** les interactions sont-elle dyadiques, c'est à dire mettent-elles en jeu deux personnes ?

- **Disponibilité :** le corpus est-il disponible ?

Corpus	Attitude	Contexte	Modalités	Dyades	Disponibilité
SEMAINE [McKeown et al., 2010]	✓	×	×	✓	✓
AMI [Carletta et al., 2006]	×	×	✓	×	✓
CID [Blache et al., 2010]	×	×	✓	✓	✓
Comparse [Giraud et al., 2013]	×	✓	✓	×	×
HuComTech [Hunyadi et al., 2012]	×	✓	✓	✓	×

TABLE 4.1: Corpus pertinents existants et critères indispensables à leur sélection.

Dans la table 4.1, nous rapportons les corpus présentés précédemment et s'ils remplissent les critères que nous avons défini. Cet inventaire nous a permis de conclure qu'aucun corpus multimodal disponible n'était tout à fait adapté à nos besoins. Ainsi, nous avons choisi de construire notre propre corpus multimodal.

4.3 ANNOTATION D'UN CORPUS MULTIMODAL D'ENTRETIENS D'EMBAUCHE

Afin d'étudier les attitudes exprimées par des recruteurs au moyen de séquences de signaux non-verbaux, nous avons réuni un corpus de vidéos d'entretiens d'embauche que nous avons annotées au niveau de l'attitude et du comportement non-verbal des recruteurs. Dans cette section, nous décrivons le corpus multimodal que nous avons collecté et son processus d'annotation.

4.3.1 Collecte

Un des partenaires du projet Tardis est la Mission Locale Val d'Oise Est (MLVOE) : les Missions Locales constituent un réseau de structures locales, implantées dans

FIGURE 4.4: Une des vidéos du corpus d'entretien d'embauches.

toute la France, dont le but est de favoriser l'insertion sociale et professionnelle de jeunes adultes[3]. Une des actions menées par la MLVOE est d'organiser des ateliers de préparation aux entretiens d'embauche. Dans ceux-ci, les personnes fournissent une offre d'emploi à laquelle ils comptent postuler, et un conseiller de la mission locale leur permet de répéter l'entretien d'embauche en jouant le rôle du recruteur. Ces entretiens sont adaptés au profil de la personne, à sa personnalité (*ex.* plutôt timide ou plutôt confiant) ainsi qu'à son expérience (*ex.* a-t-elle déjà passé des entretiens d'embauche ?).

Dans le cadre du projet Tardis, une étude a été menée à la MLVOE dans lequel certains de ces entretiens simulés ont pu être enregistrés. Nous avons ainsi pu récolter un ensemble de neuf vidéos, correspondant aux entretiens entre neuf candidats différents et cinq recruteurs. Dans cet ensemble de vidéos, nous avons choisi de conserver un premier sous-ensemble de cinq vidéos, les autres vidéos présentant des problèmes de cadrage. En revanche, les vidéos restantes, d'une durée contenue entre quinze et vingt minutes, mettent en scène les cinq différents recruteurs, et ont été enregistrées dans la même salle et avec la même configuration : le recruteur et le candidat sont assis des deux côtés d'une table, et une caméra est placée sur le côté de la scène afin de contenir les deux participants dans le cadre de la caméra (voir Figure 4.4). Nous avons choisi ce positionnement des caméras afin de déranger le moins possible les participants, quitte à limiter la visibilité des expressions faciales des participants. Enfin, pour des questions de temps, nous avons choisi de limiter l'annotation à trois vidéos, ces trois vidéos comptabilisant une durée totale de 57 minutes et 32 secondes d'interaction.

[3] `http://www.mission-locale.fr/`

4.3.2 Méthodologie d'annotation

Une fois les données collectées, nous avons étudié les différents outils d'annotation et les différentes schémas de codage disponibles afin de choisir ceux les plus adaptés à nos besoins.

4.3.2.1 Choix des outils d'annotation

Pour l'annotation du comportements non-verbal et du contexte de l'interaction, les logiciels Elan et Anvil se sont révélés adaptés à nos besoins et très proches en terme de puissance et d'ergonomie (comme nous l'avons présenté dans la section 4.1.2). Ce sont finalement des considérations pratiques qui ont orienté notre choix vers Elan : en effet Anvil présente une stabilité plus faible lors de l'utilisation de vidéos de plus de dix minutes, et Elan était utilisé par un groupe de chercheurs partenaires dans le projet Tardis, permettant ainsi le partage de ressources et d'expériences autour de l'utilisation du logiciel.

Nous avons utilisé le logiciel Praat [Boersma & Weenink, 2001] pour pré-traiter les pistes audio des entretiens d'embauches afin d'obtenir le comportement verbal des recruteurs. Praat est un logiciel libre spécialisé dans l'analyse et la reconstruction de signaux vocaux. Un de ses avantages est qu'il dispose d'un langage de scripts facile à prendre en main et très puissant permettant l'automatisation de tâches, ce qui s'est révélé utile pour traiter rapidement le signal audio des vidéos. En outre, Praat propose un format de sortie (*Textgrid*) compatible avec Elan, ce qui nous a permis ensuite de les y importer.

Enfin, pour l'annotation de l'attitude, nous avons choisi d'adopter un paradigme d'annotation continue. En effet, la représentation des attitudes d'Argyle (voir Section 2.2.2) consiste en deux dimensions continues, l'amicalité et la dominance, et se prête donc naturellement à une annotation continue. De plus, celle-ci nous permet de capturer les réactions instantanées des annotateurs, et de pouvoir comparer les différentes intensités des variations d'attitudes perçues par ces derniers. Nous avons choisi l'outil GTrace pour collecter les annotations continues d'attitude. En effet, celui-ci était disponible en open-source ce qui nous a permis de l'adapter à nos besoins, notamment d'adapter l'interface graphique, ainsi que l'échelle d'annotation et

les messages d'explication présentés à l'annotateur. La figure 4.5 présente l'échelle d'annotation de la dominance. En haut de cette interface, une description de la mesure est présentée à l'annotateur sous forme de question : *« How dominant or submissive do you perceive the recruiter ? »*.

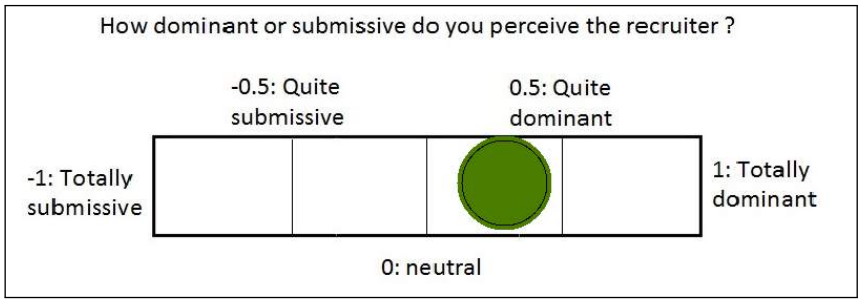


FIGURE 4.5: Échelle d'annotation présentée aux annotateurs de l'attitude des recruteurs (cas de la dominance).

4.3.2.2 Choix du schéma de codage

Comme nous l'avons présenté en section 2.2.3, le comportement verbal et non-verbal participent à l'expression d'attitudes. Cependant, le comportement d'une personne dépend fortement de la tâche qu'elle est en train de réaliser [Argyle & Cook, 1976], du type d'interaction en cours [Brunet et al., 2009] ou encore du tour de parole [Cassell & Thorisson, 1999]. Par exemple, dans une interaction dyadique, la quantité de regards dirigés vers l'interlocuteur est influencée par la tâche en cours de réalisation [Argyle & Cook, 1976] : si un objet lié à la tâche en cours est présent près de la dyade (*ex.* le *curriculum vitae* du candidat lors d'un entretien d'embauche), les deux personnes impliquées vont regarder régulièrement cet objet et moins regarder leur interlocuteur. Le rythme des changements de posture est également influencé par le tour de parole [Cassell et al., 2001]. Lors de conversations, les participants changent plus souvent de postures lorsqu'ils prennent la parole ou lorsque le sujet de conversation change.

Notre schéma de codage comprend ainsi plusieurs parties : les étiquettes concernant le *contexte de l'interaction*, le *comportement verbal*, et le *comportement non-verbal*.

Codage du contexte de l'interaction :

La table 4.2 liste les étiquettes d'annotation du contexte de l'interaction. Les étiquettes liées à la tâche en cours reflètent le type de tâches d'un entretien d'embauche. Mis à part les salutations et la fin d'un entretien, qui sont des phases particulières d'un entretien d'embauche, nous distinguons seulement si le sujet de discussion concerne un document présent dans la salle (*ex.* le CV du candidat ou l'offre d'emploi) ou non (*ex.* une question générale : « Pouvez-vous me parler de vous ? »). Enfin, une catégorie particulière d'annotation est la catégorie *Masque*, que nous avons utilisé pour des questions pratiques pour distinguer les rares instants ou le cadre d'entretien est suspendu (*ex.* lors d'une pause dans l'entretien) ou lorsque le recruteur n'est pas visible (*ex.* lors d'un ajustement de la caméra), ce qui nous a par la suite simplifié les pré-traitements des données.

Catégorie	Évènement	Étiquette
Tâche	Question liée au CV ou à l'offre d'emploi	DocumentRelated
	Question sur tout autre sujet	General
	Salutations	GreetFarewell
	Transitions	Transition
Masque	Occlusion ou pause	Mask

TABLE 4.2: Étiquettes d'annotations du contexte de l'interaction.

Codage du comportement verbal :

Pour le tour de parole, nous avons choisi de distinguer les instants où le candidat a le tour de parole, les instants où le recruteur a le tour de parole, les instants où les participants parlent simultanément (*i.e.* interruption ou recouvrement de parole) et les instants où aucun participant ne revendique le tour de parole. Afin de rendre cette annotation plus facile, nous avons d'abord procédé à une annotation du comportement verbal du recruteur et du candidat.

Pour le comportement verbal, nous avons choisi de faire la distinction entre la parole, le silence, les hésitations (vocalisations non-verbales comme « *hum* » ou « *euh* ») et les rires. Nous avons aussi défini une catégorie d'annotation pour différencier les phrases du recruteur contenant une valeur émotionnelle des phrases neutres en utilisant le schéma d'annotation de la théorie de l'*Interaction Process Analysis* de Bales [Bales, 1950]. L'annotation de ces phrases a cependant montré que le corpus ne comptait

que très peu de phrases comportant une valeur émotionnelle. La table 4.3 liste les étiquettes d'annotation du comportement verbal des recruteurs.

Catégorie	Évènement	Étiquette
Vocale	Parole	VSpeaking
	Silence	VSilent
	Rire	VLaughing
	Hésitation	VFilledPause
Verbale	Phrase liée à la tâche	TO
	Phrase socio-émotionnelle positive	SEPos
	Phrase socio-émotionnelle négative	SENeg
Tour de parole	Recruteur	Recruiter
	Candidat	Candidate
	Interruption : les deux participants parlent simultanément	Interruption
	Silence prolongé : personne n'a la parole	Silence

TABLE 4.3: Étiquettes d'annotation du comportement verbal.

Codage du comportement non-verbal :

Le schéma d'annotation des comportements corporels de MUMIN est assez détaillé : toutefois, celui-ci n'est pas complètement adapté à nos besoins. En effet, on peut remarquer que la posture, les directions de regards et les directions de tête ne sont pas prises en compte (*ex.* les hochements de tête sont inclus, mais pas le fait de garder la tête haute pendant un moment). En outre, certaines expressions faciales assez détaillées auraient été difficiles à annoter compte tenu du point de vue adopté dans les vidéos du corpus. Nous avons donc décidé d'adapter le schéma MUMIN à nos besoins.

La table 4.4 liste les étiquettes d'annotation du comportement non-verbal du recruteur : le regard, les mouvements et positions de tête sont inclus, car ils sont liés à l'expression d'attitudes (voir Section 2.2). Nous n'avons pas essayé d'annoter les expressions faciales à un niveau très précis (comme les *action units* d'Ekman [Ekman & Friesen, 1977]) à cause de la distance entre la dyade et la caméra, cependant nous avons inclus les mouvements de sourcils (froncements, haussements) et les sourires. Les postures et les gestes, deux comportements essentiels dans l'expression de l'attitude (voir Section 2.2), sont inclus. Pour les gestes, nous considérons aussi les manipulations d'objets et les gestes *adaptors* : en effet, les personnes jouant avec un stylo, ou se grattant, sont perçues comme nerveuses [Carney et al., 2005]. Enfin,

nous avons aussi ajouté différentes positions de repos des mains : bras croisés, mains jointes ou non.

Catégorie	Expression	Étiquette	Optionnel
Regard	Vers l'interlocuteur	GazeAt	
	Vers un objet (*ex.* CV)	GazeObject	
	Vers le haut	GazeUp	
	Vers le bas	GazeDown	
	Vers le coté	GazeSide	
Tête (direction)	Vers l'interlocuteur	HeadAt	
	Vers le haut	HeadUp	Int[1]
	Vers le bas	HeadDown	Int[1]
	Vers le côté	HeadSide	Int[1]
	Penchée sur le côté	HeadTilt	Int[1]
Tête (mouvement)	Hochement vertical	HeadNod	Int[1], Rep[2]
	Hochement horizontal	HeadShake	Int[1], Rep[2]
Sourcils	Haussement	EyebrowUp	Int[1]
	Froncement	EyebrowDown	Int[1]
Bouche	Sourire	Smile	Int[1]
Posture	Droit sur la chaise	BodyStraight	
	Penchée vers l'interlocuteur	BodyLean	Int[1]
	Penchée en arrière	BodyRecline	Int[1]
Gestes	Gestes communicatifs	GestComm	Int[1], Spa[3]
	Manipulation	GestObjManip	
	Adaptor (*ex.* se gratter)	GestAdaptor	BPart[4]
Positions des mains	Posées sur la table (séparées)	RestOver	
	Sous la table	RestUnder	
	Bras croisés	RestArmsCrossed	
	Jointes	RestHandsTogether	

TABLE 4.4: Étiquettes d'annotation du comportement non-verbal.

Des paramètres optionnels permettent de préciser l'expressivité de certains signaux non-verbaux :

- $Int \in \{High, Normal, Low\}$ est un paramètre optionnel d'intensité, de faible (Low) à élevée ($High$). Il permet de préciser si une expression faciale ou un geste est particulièrement intense (*ex.* sourire très prononcé, geste puissant).
- $Rep \in \{Yes, No\}$ est un paramètre optionnel utilisé pour les mouvements de tête, afin de distinguer des mouvements isolés ou répétés de manière continue (*ex.* un hochement de tête, ou plusieurs successifs).

- $Spa \in \{Small, Normal, Large\}$ est un paramètre optionnel d'amplitude spatiale utilisé pour préciser si un geste est particulièrement ample ou très petit.
- $BPart \in \{Face, Hair, Neck, Hands, Body, Other\}$ est un paramètre optionnel pour définir la partie du corps qui est touchée par le participant lors de gestes de la catégorie des *Adaptors* (*ex.* la tête, les bras, *etc.*).

4.3.3 Processus d'annotation

Pour l'annotation du comportement verbal, nous avons d'abord programmé des scripts Praat afin d'obtenir rapidement une première segmentation du signal audio des entretiens d'embauche du corpus, entre les instants où une personne parle et les instants où personne ne parle plus. Nous avons ensuite segmenté ces annotations pour différencier quelle personne était en train de parler : on obtient alors une trame comportant les instants où le recruteur parle ou non, et une trame comportant les instants où le candidat parle ou non. Une fois ces deux trames obtenues, le tour de parole a été annoté sous Elan, parallèlement à l'annotation du comportement non-verbal, de manière à ce que celui-ci soit aligné sur le signal audio.

Modalité	Nombre d'annotations	Score κ
Regard	836	0.95
Tete (direction)	658	0.75
Gestes	313	0.80
Tete (mouvements)	281	0.71
Positions de main	245	0.80
Sourcils	156	0.62
Postures	123	0.93
Sourires	91	0.78

TABLE 4.5: Résultats du processus d'annotation du comportement non-verbal.

Un annotateur unique a annoté complètement le comportement non-verbal des recruteurs et du contexte de l'interaction des trois vidéos. Nous n'avons annoté que le comportement des recruteurs car notre objectif était de modéliser le comportement d'un Agent Conversationnel Animé lors de son discours, et le comportement des candidats n'était alors pas pertinent. Un mois après le processus d'annotation, une seconde annotation a été réalisée par le même annotateur sur une sous-partie du corpus. Nous avons calculé le score kappa κ pour les différentes modalités du

comportement non-verbal afin d'évaluer la validité des annotations (voir Table 4.5). Celui-ci est satisfaisant pour toutes les modalités ($\kappa > 0.70$), hormis pour les mouvements de sourcils ($\kappa = 0.62$), dont le score plus faible peut être expliqué par la distance entre la caméra et les participants, peut-être trop importante pour permettre une détection précise. Le nombres d'étiquettes annotées par modalité et les scores de *kappa* correspondants sont présentés en table 4.5.

Douze personnes ont pris part à l'annotation de l'attitude des recruteurs. Pour réduire la difficulté de la tâche d'annotation, chaque annotateur avait pour consigne d'annoter une seule dimension à la fois, comme le recommandent Cowie *et al.* [Cowie et al., 2012]. Comme les vidéos sont longues, les annotateurs avaient l'option de mettre la vidéo en pause à tout moment. De plus, nous avons filtré l'audio des vidéos pour le rendre inintelligible, par le biais du logiciel Praat. En effet, notre objectif étant de modéliser l'expression d'attitudes par le comportement non-verbal, nous ne voulions pas que les annotateurs soient influencés par le discours du recruteur dans leur jugement de son attitude. Au total, nous avons récupéré de deux à trois traces d'annotation par vidéo et par dimension.

Si l'évaluation de l'accord inter-annotateur est aisée pour des données discrètes (deux annotateurs sont d'accord ou non sur l'étiquette qu'ils choisissent pour décrire une donnée), l'évaluation de l'accord inter-annotateur est moins évidente dans le cas d'annotations continues [Metallinou & Narayanan, 2013]. En effet, on peut s'attendre, du fait de l'échelle continue et de la nature continue de l'annotation, à ce que les annotateurs ne soient pas exactement d'accord sur la valeur de la dimension annotée à un instant donné [Cowie & McKeown, 2010]. De plus, certains auteurs [Vaasen et al., 2012] avancent qu'il est fondamentalement difficile d'obtenir un fort accord inter-annotateur au sujet de phénomènes subjectifs et complexes, comme l'attitude exprimée par une personne, et que cela n'est peut-être pas essentiel dans un contexte de génération de comportement : *« Since the goal of the application is to simulate human behaviour, these results also imply that it is not critical for the final application to reach a perfect level of prediction. In fact, due to the subjective nature of the annotation process, an objectively "correct" does not exist »*.

Cowie *et al.* suggèrent que pour réduire les problèmes de bruitage et d'échelle spécifiques aux traces d'annotation continue, on peut procéder à une stylisation des données, c'est à dire à considérer les variations de la courbe plutôt que ses valeurs

absolues [Cowie & McKeown, 2010]. Par exemple, les annotateurs ne seront peut-être pas d'accord sur les valeurs exactes de la courbe, mais peut-être sur le fait que la courbe augmente, diminue, ou reste stable. Ces auteurs conseillent aussi de faire la moyenne les valeurs des traces sur des intervalles courts, d'une longueur recommandée entre une et trois secondes, afin d'effacer le bruit inhérent à ce type d'annotation.

Pour analyser la fiabilité des données annotées d'attitude, nous avons ainsi d'abord calculé, pour chacune des traces d'attitudes, une valeur moyenne sur des intervalles de trois secondes (l'attitude évoluant plus lentement que les émotions, nous avons pris la valeur la plus haute recommandée par [Cowie & McKeown, 2010]). Nous avons ensuite transformé ces données en variations, c'est à dire que plutôt que de comparer les valeurs absolues moyennes d'une trace à l'autre, nous calculons la variation entre les intervalles successifs pour chaque trace, et ce sont ces variations que nous comparons. Sur l'ensemble des vidéos, l'alpha α de Cronbach a révélé une fiabilité moyenne ($\alpha = 0.489$), avec la meilleure vidéo présentant un score de $\alpha = 0.646$. Ces valeurs seraient considérées plutôt faibles dans un processus d'annotation classique : toutefois, comme nous l'avons indiqué plus haut, l'annotation continue en temps et en valeur est susceptible de produire des valeurs de fiabilité plus faibles que pour des annotations d'évènements ou d'intervalles [Cowie & McKeown, 2010], et l'annotation de l'attitude a aussi tendance à être source de désaccord [Vaasen et al., 2012].

4.3.4 Analyse des traces d'attitudes

Afin de réduire le bruit inhérent aux données obtenues par annotation continue, nous avons procédé à un stylisation des traces obtenues (voir Section 4.3.3). Pour cela, nous avons d'abord segmenté les données en plateaux (les segments où les valeurs d'attitude sont stables) et en pentes (les segments où l'attitude varie). Ensuite, nous avons lissé les traces en supprimant les plateaux très courts (inférieurs à une seconde), en regroupant les pentes avoisinantes, et en supprimant les pentes très légères (variations inférieures à un vingtième de l'échelle d'annotation). Cette étape est représentée sur la figure 4.6.

Après le processus d'annotation et de stylisation, nous comptabilisons 120 plateaux

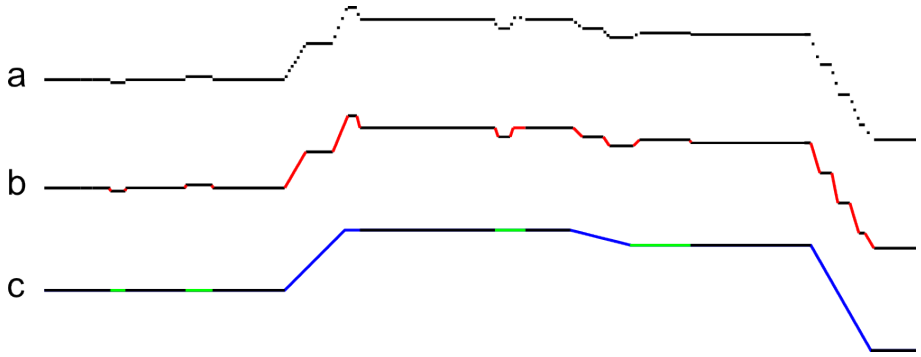


FIGURE 4.6: Stylisation de traces d'attitudes. (a) Trace brute, représentant le résultat de l'annotation d'une vidéo par un annotateur. (b) Détection des segments où l'attitude varie (pentes : rouge). (c) lissage des pentes trop courtes (bleu) et des plateaux trop courts (vert).

de dominance, 117 variations de dominance, 119 plateaux d'amicalité et 116 variations d'amicalité. Les durées et valeurs moyennes des pentes et plateaux par vidéos sont présentées dans la table 4.6. Des différences importantes peuvent être observées entre les vidéos : à partir des valeurs moyennes des plateaux, on voit que le premier recruteur est généralement perçu comme assez dominant et le deuxième recruteur est perçu comme assez inamical. Les plateaux et pentes semblent avoir une durée plus longue dans le cas de la dominance que dans le cas de l'amicalité, ce qui pourrait suggérer que cette dimension est moins sujette aux changements soudains.

Vidéo	Durée	Dimension	Plateaux			Pentes		
			Nombre	Valeur moyenne [-1, 1]	Durée moyenne	Nombre	Valeur moyenne [0, 1]	Durée moyenne
1	16 :26	Dominance	41	0.41	7.9s	40	0.21	7.6s
		Amicalité	46	0.08	7,1s	45	0.25	5.2s
2	17 :53	Dominance	28	0,19	20,4s	27	0.19	11.9s
		Amicalité	26	-0.33	11.3s	25	0.22	6.5s
3	23 :13	Dominance	51	0.13	9.8s	50	0.24	9.2s
		Amicalité	47	0.06	6.6s	46	0.29	8.7s

TABLE 4.6: Plateaux et pentes observés dans les traces d'attitude pour les trois vidéos annotées.

CONCLUSION

Dans ce chapitre, nous avons présenté un aperçu des techniques et des outils pour la création et l'annotation de corpus multimodaux. Nous avons ensuite présenté une sélection de corpus multimodaux existants, et avons constaté qu'aucun corpus disponible ne remplissait tous les critères indispensables à notre travail. Nous avons ainsi décidé d'utiliser des enregistrements d'entretiens d'embauche collectés dans le cadre du projet Tardis. Nous avons réuni neuf enregistrements d'entretiens d'embauche. Nous avons adapté un schéma de codage à notre objectif de recherche, et avons procédé à l'annotation du comportement non-verbal du recruteur, du contexte de l'interaction et de l'attitude du recruteur dans trois de ces enregistrements.

Lors du processus d'annotation, nous avons collecté plus de huit mille étiquettes d'annotation sur le comportement verbal, le comportement non-verbal et le contexte de l'interaction. Nous avons aussi récolté des traces d'annotation d'attitudes pour les vidéos considérées. Nous avons utilisé une technique de stylisation des données afin de réduire le bruit inhérent aux annotations continues.

Dans le chapitre suivant, nous présentons une méthodologie permettant d'extraire de ce corpus des informations sur la manière dont les signaux non-verbaux sont perçus en terme d'attitudes. Ces informations seront ensuite utilisées pour déterminer comment un Agent Conversationnel Animé peut exprimer une attitude par son comportement non-verbal. Pour extraire ces informations, nous avons utilisé une technique de *fouille de motifs séquentiels*, qui nous a permis d'obtenir un ensemble de séquences de signaux non-verbaux fréquemment observées avant différents types de variations d'attitudes.

Synthèse du chapitre

1. Les corpus multimodaux sont une ressource essentielle pour la conception et la modélisation d'ACAs. Lors de la création d'un corpus multimodal, plusieurs aspects essentiels doivent être considérés sur la manière de collecter les données. De nombreux outils sont utiles pour annoter les données.
2. De nombreux corpus multimodaux, réalisés dans des contextes de recherche particuliers, peuvent être réutilisés pour étudier de nouvelles questions ou alimenter d'autres modèles. Cependant, il n'existe pas de corpus tout à fait adapté à notre problématique.
3. Nous avons adapté un schéma de codage existant pour annoter notre corpus. Nous avons obtenu plus de huit mille étiquettes d'annotation et au moins deux traces d'annotations d'attitudes par dimension par vidéo.
4. Une analyse des traces d'attitudes récoltées montre que les recruteurs des différentes vidéos expriment globalement des attitudes diverses.

5

Fouille de séquences de signaux non-verbaux

Comme nous l'avons présenté dans la section 2.3, un signal non-verbal peut être interprété selon plusieurs perspectives. En particulier, l'interprétation d'un signal non-verbal peut être modifiée par des signaux proches [Keltner, 1995, With & Kaiser, 2011, Jack et al., 2014]. Lors de la planification de comportement d'un Agent Conversationnel Animé, il est donc nécessaire de considérer des séquences de signaux non-verbaux, et pas uniquement les signaux non-verbaux indépendamment les uns des autres. En outre, nous avons retenu de la définition des attitudes interpersonnelles de Scherer [Scherer, 2005] que celles-ci ne s'expriment pas de manière prototypique à un instant donné mais plutôt qu'elles affectent le comportement d'une personne sur de longues périodes. Nous avons donc proposé de considérer des séquences de signaux non-verbaux plutôt que des signaux non-verbaux indépendants lors de la planification de comportement pour l'expression d'attitudes.

Une large littérature existe sur l'influence de signaux non-verbaux étudiés de manière indépendante sur l'expression d'attitudes (voir Section 2.2.3), mais celle-ci ne nous fournit que peu d'informations sur le lien entre des séquences de signaux non-verbaux et l'expression d'attitudes. Nous avons donc décidé d'extraire ces informations automatiquement à partir d'un corpus multimodal.

Dans ce chapitre, après avoir présenté différentes techniques pour l'analyse ou la modélisation de séquences d'évènements, nous introduisons une méthodologie pour l'extraction de séquences de signaux non-verbaux que nous appliquons à l'étude de séquences caractéristiques de l'expression d'attitudes sociales. Ces séquences forment la base du modèle de planification de comportement pour l'expression d'attitudes qui sera présenté au chapitre 6.

5.1 Techniques d'analyse séquentielle

Dans cette section, nous présentons tout d'abord un aperçu de différentes techniques permettant d'analyser des données séquentielles.

5.1.1 Analyse séquentielle

L'analyse séquentielle fait référence à un ensemble de techniques statistiques pour la description et l'analyse de données séquentielles [Bakeman & Quera, 2011]. Ces techniques permettent d'obtenir des tables de probabilités de transitions entre différents types d'évènements contenus dans les données étudiées. L'analyse séquentielle *par décalage* (*sequential lag analysis*) permet de calculer les probabilités conditionnelles que différents types d'évènements arrivent après un premier évènement, appelé l'*antécédent* [Sackett, 1987, Bakeman & Quera, 2011]. Pour analyser des évènements d'une trame survenant directement après l'antécédent, on parle de *lag 1 analysis*, de *lag 2 analysis* pour les évènements arrivant en deuxième position, et ainsi de suite.

Un des premiers domaines de recherche qui a donné lieu à la création des techniques d'analyse séquentielle est l'étude de la participation d'enfants à des activités de groupe : Bakeman et Brownlee calculent la probabilité que des enfants passent d'un type d'activité à un autre à partir de données récupérées sur le terrain [Bakeman & Brownlee, 1980]. Ils montrent que les enfants commencent par jouer seuls, puis passent par une étape de jeu en parallèle aux autres enfants avant d'évoluer vers une activité de jeu en groupe. L'analyse séquentielle par décalage a aussi été utilisée dans le domaine de l'analyse du comportement non-verbal. Allwood *et al.* déterminent les signaux non-verbaux précédant et suivant des *feedbacks* verbaux (vocalisations réalisées par un interlocuteur, *ex.* « *hmm-hmm* » ou simplement « *oui* ») [Allwood et al., 2007b]. L'utilisation de l'analyse séquentielle par décalage leur permet de calculer les probabilités conditionnelles qu'un *feedback* verbal soit précédé ou suivi de certains signaux non-verbaux. Par exemple, ils observent qu'un *feedback* verbal est régulièrement précédé par une orientation du corps vers l'interlocuteur (*Body Towards*, $p = 0.2857$) ou d'un hochement de tête (*Single Nod*, $p = 0.2222$).

5.1.2 Détection de t-patterns

Un formalisme intéressant pour la modélisation de motifs d'évènements est celui des *t-patterns* proposé par Magnusson [Magnusson, 2000]. Les *t-patterns* sont des motifs hiérarchiques d'évènements possédant des relations temporelles *relativement invariantes*. Un exemple de *t-pattern* est présenté dans la figure 5.1. Le logiciel THEME est spécialisé dans la détection de ces motifs.

La notion de relations temporelles *relativement invariantes* est centrale dans le concept de *t-pattern*. Pour déterminer si une relation temporelle entre des évènements d'un *t-pattern* est relativement invariante, une méthode statistique est utilisée. A partir des distributions observées des évènements étudiés, et en considérant que ces évènements sont indépendants, on calcule le nombre théorique de motifs dans lesquels la relation temporelle considérée aurait pu arriver par chance. Si la relation temporelle étudiée est observée plus souvent dans les données que ce nombre théorique, alors on considère que la relation temporelle est relativement invariante.

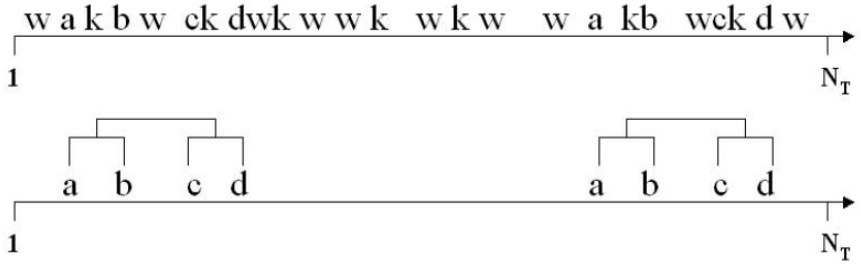


FIGURE 5.1: Exemple de *t-pattern*, où le motif A-B-C-D, difficile à repérer à l'œil nu, est extrait d'une chaine de caractères (figure reproduite de [Jonsson et al., 2010]).

La détection de *t-patterns* a été appliquée à l'étude de nombreux domaines, comme la détection de motifs d'évènements récurrents dans le football ou encore dans les interactions homme-chien [Jonsson, 2011]. Ces techniques permettent de détecter des motifs d'évènements qui seraient difficiles à repérer par des humains [Jonsson et al., 2010].

5.1.3 Modèles de Markov

Les modèles de Markov sont des outils utilisés pour modéliser des processus stochastiques vérifiant la propriété de Markov. L'évolution d'un système vérifiant la propriété de Markov ne dépend que de l'état dans lequel ce système se trouve au moment présent, et pas des états passés. Plus précisément, si X_n représente l'état d'un système à l'étape n, et que $i_0, ..., i_{n-1}, i_n, j$ sont une suite d'états, alors : $P(X_{n+1} = j | X_0 = i_0, ..., X_{n-1} = i_{n-1}, X_n = i_n) = P(X_{n+1} = j | X_n = i_n)$. Les chaînes de Markov sont le type de modèle de Markov le plus simple. Dans ce cas, le modèle correspond à un automate à états finis dans lequel les transitions entre états sont représentées par des probabilités.

Les chaînes de Markov cachées (ou *Hidden Markov Models*, abrégées en HMM) sont une extension des chaînes de Markov. La différence majeure est que l'état du système à un instant t n'est pas directement observable, mais que chaque état produit des observations auxquelles on peut accéder (voir Figure 5.2). Les observations produites dans chaque état dépendent d'une autre distribution de probabilité [Rabiner, 1990].

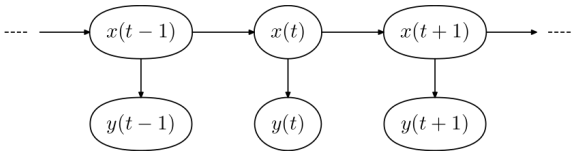


FIGURE 5.2: Représentation d'un modèle de Markov caché, où $x(t)$ est l'état (non-observable) du système à l'instant t et $y(t)$ l'observation produite par l'état $x(t)$.

Les modèles de Markov cachés ont été appliqués avec succès à des problèmes variés comme la reconnaissance de parole [Rabiner, 1990] ou le traitement du langage naturel [Charniak et al., 1993]. Dans le domaine de la modélisation du comportement non-verbal, Lee et Marsella ont utilisé des modèles de Markov cachés pour prédire à quels moments un Agent Conversationnel Animé en train de parler doit hocher la tête, par rapport aux propriétés linguistiques de la phrase à prononcer [Lee & Marsella, 2010].

5.1.4 Fouille de motifs séquentiels

La fouille de motifs séquentiels (*sequential pattern mining*) est un autre ensemble de techniques dont le but est l'extraction de motifs fréquents dans des données de nature séquentielle. Un des intérêts de ces techniques est que les séquences extraites ne sont pas forcément constituées d'évènements contigus. Ainsi, des séquences d'évènements régulières pourront être identifiées même si d'autres évènements, indépendants, sont parfois observés au milieu de cette séquence.

Le problème de la fouille de motifs séquentiels peut être considéré comme un cas particulier du problème de fouille de règles d'associations (*association rule learning*) [Masseglia et al., 2004]. Dans les deux cas, l'objectif est d'extraire des connaissances sur les types d'évènements arrivant fréquemment ensemble dans des données. Un exemple classique de problème de fouille de règles d'association est la recherche des produits que des clients achètent ensemble au supermarché (*ex.* la règle $Pain \rightarrow Beurre$ indique *si un client achète du pain, alors il achète aussi du beurre*). Les règles d'associations peuvent être caractérisées par différentes mesures. Le *support* d'un évènement indique sa fréquence d'apparition dans les données. La *confiance* d'une règle indique la fréquence à laquelle celle-ci se vérifie.

Client	Mercredi	Jeudi	Vendredi
1	Pain, Œufs	Beurre	Tomates
2	Pain	N/A	Œufs, Chocolat
3	N/A	Pain	Beurre, Vin
4	Œufs	Tomates	Beurre
5	Chocolat	Œufs	Vin

TABLE 5.1: Exemple de données utilisés dans la fouille de motifs séquentiels (problème de l'identification de produits achetés lors de jours successifs). Une règle que l'on pourrait extraire de ces données est qu'un client achetant du pain a deux chances sur trois d'acheter du beurre dans les deux jours. Cette règle se note $Pain \rightarrow Beurre$, possède un $Support$ de 3/5 (du pain est acheté trois fois dans les données), et une $Confiance$ de 2/3 (sur les trois clients ayant acheté du pain, deux ont acheté du beurre dans les deux jours).

La différence entre la fouille de motifs séquentiels et la fouille de règles d'associations est que dans le cas de la fouille de motifs séquentiels, on cherche à identifier un ordre dans lequel les évènements arrivent. Pour reprendre le même exemple, on pourrait rechercher des régularités dans les achats réalisés par un client lors de jours

successifs (*ex. si un client achète du pain, alors il achètera du beurre le lendemain.* Nous présentons un exemple de jeu de données et d'un motif séquentiel dans la table 5.1. Ce type de techniques a été utilisé pour détecter des motifs intéressants dans le cadre de l'analyse des protéines de l'ADN [Ferreira & Azevedo, 2005], ou encore dans l'analyse des affects de personnes jouant à des jeux vidéo (*ex.* frustration) par rapport aux séquences de touches qu'ils pressent [Martínez & Yannakakis, 2011].

5.1.5 Conclusion

Les techniques d'analyse séquentielle présentent l'avantage d'être relativement simples à réaliser et de fournir des résultats simples à interpréter. Un inconvénient majeur de ces différentes méthodes est que seules des paires d'évènements sont analysées, or il est possible qu'une séquence de signaux non-verbaux de plus de deux signaux soit pertinente pour l'expression d'attitudes.

Le formalisme des *t-patterns* est intéressant car il permet d'extraire des motifs d'évènements aux relations temporelles complexes. Toutefois, dans notre contexte de génération de séquences de signaux non-verbaux alignées sur un texte, il pourrait être difficile de réutiliser des motifs comprenant de telles contraintes. Par exemple, l'utilisation d'un logiciel de synthèse de parole implique que nous n'avons pas le contrôle sur la durée d'énonciation des mots par l'agent. Or, les gestes de l'agent doivent être synchronisés avec le texte [McNeill, 1992] : il se pourrait donc qu'aucun des *t-patterns* extraits de notre corpus multimodal ne puisse s'adapter à une nouvelle phrase.

Les modèles de Markov cachés présentent aussi leurs avantages. Cependant, l'hypothèse de Markov ne se vérifie pas dans notre domaine : par exemple, l'expression de l'embarras décrite par Keltner implique trois signaux successifs [Keltner, 1995]. Nous devrions donc rejeter l'hypothèse de Markov (*ex.* considérer que l'état d'un système à l'instant $n+1$ dépend pas que de l'état à l'instant n, mais aussi de l'état à l'instant $n-1$, $n-2$, ...) , ce qui impliquerait de devoir modéliser les probabilités d'occurrences de toutes les séquences possibles de plus de deux signaux non-verbaux. Or, celles-ci n'apparaissent peut-être pas toutes dans notre corpus multimodal. De plus, les modèles de Markov cachés ne permettent pas de modéliser le fait que, au milieu d'une séquence de signaux non-verbaux, d'autres signaux totalement indépendants

de cette séquence peuvent être observés.

Notre objectif est d'extraire des séquences de signaux non-verbaux qui soient caractéristiques de l'expression d'attitudes. Nous ne voulons pas mettre de restrictions sur la longueur des séquences : les techniques d'analyse séquentielle ne sont donc pas applicables à notre problème. De plus, au cours d'une séquence de signaux non-verbaux intéressante, il se peut que d'autres signaux soient produits sur d'autres modalités, du fait de la multiplicité des modalités et des fonctions du comportement non-verbal. Les modèles de Markov cachés ne sont donc pas adaptés à notre problème. Enfin, gardant à l'esprit que nous voulons utiliser les séquences extraites dans un but de planification de comportement, nous ne choisissons pas l'extraction de t-patterns : les relations temporelles complexes entre les signaux des motifs extraits rendraient leur utilisation difficile dans un contexte de planification. En conclusion, nous nous orientons vers les techniques de fouille de motifs séquentiels, qui répondent à nos besoins.

5.2 FOUILLE DE SÉQUENCES DE SIGNAUX NON-VERBAUX

Dans cette section, nous présentons une méthodologie pour l'extraction de séquences de signaux non-verbaux exprimant différents types de variations d'attitude. Nous appliquons cette méthodologie à notre corpus multimodal constitué de vidéos d'entretiens d'embauche (voir Chapitre 4), et présentons les résultats obtenus.

La méthodologie que nous proposons est composée des étapes suivantes :

1. Les annotations de signaux non-verbaux sont pré-traitées : les segments non considérés sont retirés (*i.e.* quand le recruteur n'est pas visible), et les signaux non-verbaux sont regroupés sur une trame unique et stockés sous la forme d'une (longue) séquence.

2. Nous identifions les intervalles où les attitudes annotées varient et les intervalles où les attitudes sont stables.

3. Une technique de partitionnement de données (*clustering*) est appliquée pour regrouper les évènements de variations d'attitude selon l'amplitude de la variation.

4. Les trames de signaux non-verbaux sont segmentées par rapport aux instants

où les attitudes varient, et les segments obtenus sont regroupés en fonction du type de variation qu'ils précèdent.

5. Un algorithme d'extraction de motifs séquentiels est appliqué sur chaque groupe de segments séparément, ce qui nous permet d'obtenir un ensemble de séquences fréquentes pour chaque type de variation d'attitude.

6. Des mesures de qualité sont calculées pour caractériser les séquences fréquentes extraites de chaque partition.

Dans les prochaines sections, nous présentons en détail les étapes de cette méthodologie.

5.2.1 Pré-traitement des trames de signaux non-verbaux

La première étape consiste à pré-traiter les données de signaux non-verbaux. Celles-ci sont codées dans un format XML spécifique au logiciel Elan. Nous avons écrit un programme qui interprète ces fichiers et regroupe les annotations sur une trame unique, créant ainsi une grande séquence contenant toutes les annotations de signaux non-verbaux d'une interaction ordonnées temporellement (voir Figure 5.3). Seul l'instant où un signal a été déclenché est conservé dans cette sequence résultante. Cela nous permet de simplifier l'extraction et la planification de séquences en ne considérant que l'ordre entre plusieurs signaux. Nous perdons cependant de l'information sur la durée des signaux ou sur leur instant de fin. Nous discutons cette limite de notre travail dans la section 9.2.2.

Nous segmentons ensuite cette grande séquence en fonction du contexte d'étude. Par exemple, le comportement d'une personne est largement différent dans le cas où elle est en train de parler et dans le cas où elle prend la parole. Dans notre cas, nous retirons les instants où le recruteur ne parle pas, car nous étudions l'expression de l'attitude lors du discours. Nous retirons aussi les instants où le recruteur n'est pas totalement visible (*i.e.* une annotation « *Mask* » est présente).

Dans cette étape, nous regroupons aussi les signaux non-verbaux déclenchés au même moment en un seul élément (*ex.* un sourire *Smile* et un hochement de tête *HeadNod* simultanés sont regroupés en un évènement unique $< Smile, HeadNod >$). Pour cela, nous alignons les instants de début des annotations des signaux non-

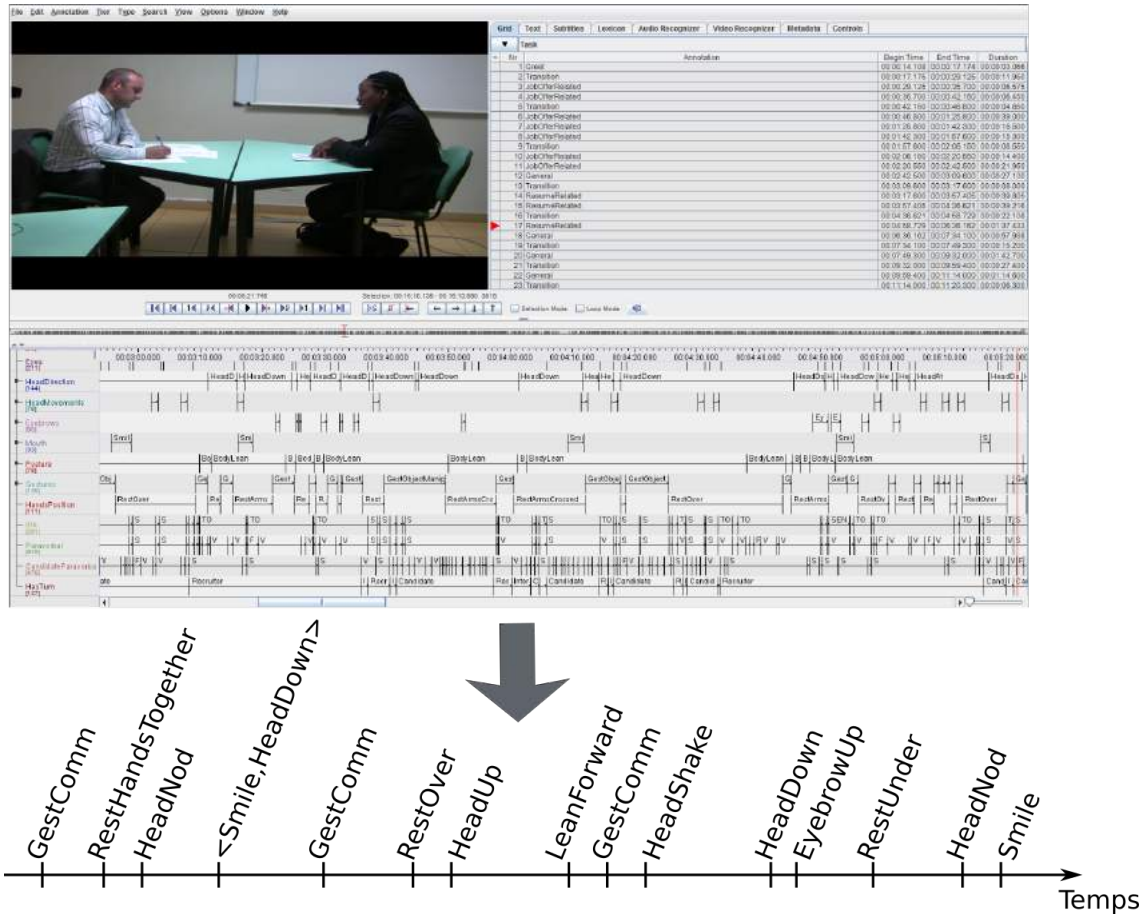


FIGURE 5.3: Transformation des fichiers d'annotations en trames de signaux non-verbaux. Seul l'instant de départ d'un signal est conservé. Nous pouvons ainsi établir une grande séquence de signaux non-verbaux représentant l'ordre dans lequel les signaux du recruteur sont réalisés dans chaque vidéo.

verbaux au quart de seconde près (*ex.* un sourire ayant été annoté entre $3,21s$ et $4,91s$ sera aligné entre $3,25s$ et $5s$). Cet intervalle d'un quart de seconde à été choisi de manière heuristique. Si plusieurs évènements sont déclenchés au même instant après cet alignement, alors ils sont regroupés en un évènement unique dans la trame de signaux non-verbaux.

5.2.2 Identification des intervalles de variation d'attitude

L'objectif de la deuxième étape consiste à identifier les intervalles où l'attitude du recruteur varie et les intervalles où l'attitude est stable. Ces informations seront

ensuite utilisées afin de segmenter les trames de signaux non-verbaux. Nous avons présenté la technique que nous utilisons dans la section 4.3.4. Les traces d'attitudes sont lissées, et nous obtenons ainsi des intervalles où l'attitude est stable, et des intervalles où l'attitude varie.

5.2.3 Partitionnement des variations d'attitude

Les variations d'attitudes de notre corpus présentent un large éventail de valeurs : en effet les annotateurs déplaçaient parfois le curseur d'annotation de manière rapide, ce qui pourrait indiquer la perception d'un fort changement dans l'attitude affichée par le recruteur, alors que parfois les déplacements du curseur étaient plus légers. Nous avons utilisé une technique de partitionnement des données (*clustering*) pour regrouper automatiquement les variations d'attitudes en des groupes d'intensité similaire. Afin de différencier les faibles et fortes diminutions et augmentations d'attitude, nous avons appliqué l'algorithme de partitionnement des K-moyennes (*K-means clustering*) avec une valeur de $K = 4$. Ce procédé est représenté dans la figure 5.4.

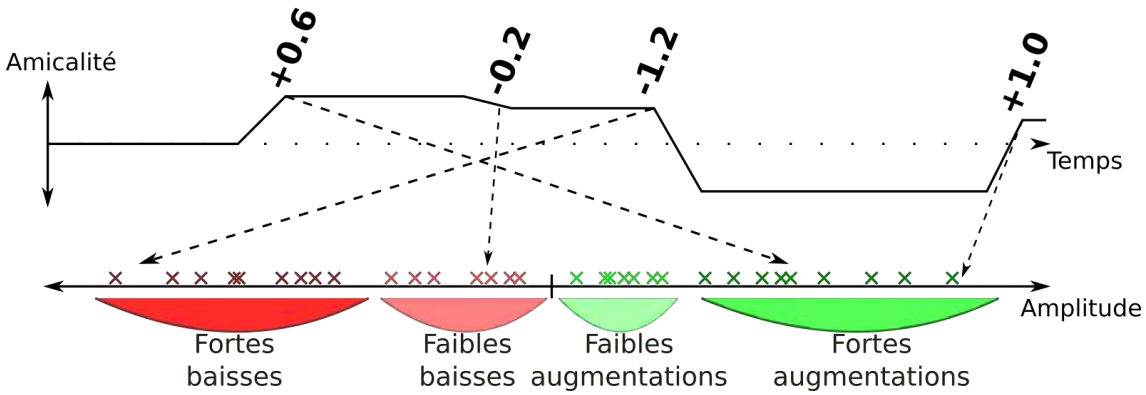


FIGURE 5.4: Identification des instants de variations d'attitude et partitionnement par amplitude de variation. L'algorithme *K-means* divise automatiquement les variations d'attitude en K partitions en fonction de leurs amplitudes de variation.

5.2.4 Segmentation et regroupement des données par type de variation d'attitude

Une fois les intervalles de variation d'attitude identifiés et regroupés en fonction de leur intensité, nous les utilisons pour segmenter les trames de signaux non-verbaux. Lors de cette étape, nous avons dû choisir l'instant précis de segmentation. Nous pouvions segmenter les données au début des variations d'attitude (*i.e.* quand l'annotateur commence à déplacer le curseur), à la fin des variations d'attitude (*i.e.* l'attitude est à nouveau stable, l'annotateur a arrêté de déplacer le curseur), ou à n'importe quel moment entre ces deux instants. Nous avons choisi de segmenter les trames de signaux non-verbaux aux instants où l'attitude a fini de varier (voir Figure 5.5). Nous appelons les instants où les attitudes cessent les *instants de variation d'attitude*.

Nous regroupons les segments obtenus en fonction du type de variation d'attitude qu'ils précèdent. Une hypothèse centrale de notre travail est que les variations d'attitude perçues par les annotateurs sont déclenchées par les séquences de signaux non-verbaux qui précèdent ces variations. Par exemple, sur la figure 5.5, la première augmentation d'amicalité (+0.6) est provoquée par la séquence de signaux $GestComm \rightarrow RestUnder \rightarrow HeadNod \rightarrow Smile$. En utilisant cette procédure, nous obtenons 219 séquences de signaux non-verbaux ayant déclenché des variations de dominance et 247 ayant déclenché des variations d'amicalité (voir Table 5.2).

5.2.5 Fouille de motifs séquentiels

La cinquième étape consiste à appliquer une méthode de fouille de motifs séquentiels séparément sur chaque groupe de séquences de signaux non-verbaux. Nous avons utilisé l'algorithme Generalized Sequential Patterns (GSP) décrit dans [Srikant & Agrawal, 1996]. L'algorithme GSP requiert en entrée une valeur de support minimal Sup_{min}. Ce nombre indique qu'une séquence doit être présente au minimum Sup_{min} fois pour être considérée comme fréquente. La sortie de l'algorithme est l'ensemble des séquences de signaux trouvées dans les segments analysés qui sont présentes au moins Sup_{min} fois. Par exemple, en utilisant $Sup_{min} = 3$, chaque séquence de signaux présente au moins 3 fois dans les données sera extraite.

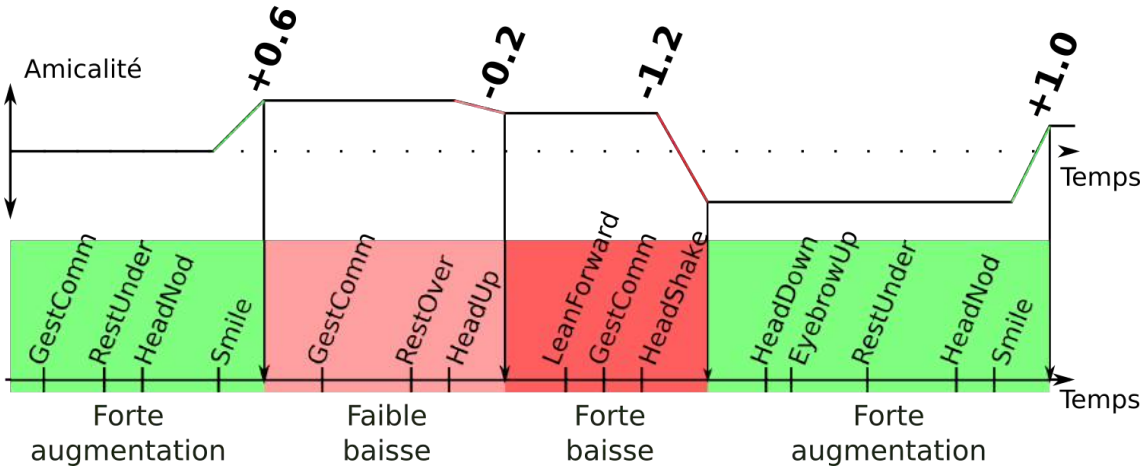


FIGURE 5.5: Segmentation des trames de signaux non-verbaux en séquences regroupées par type de variation d'attitude.

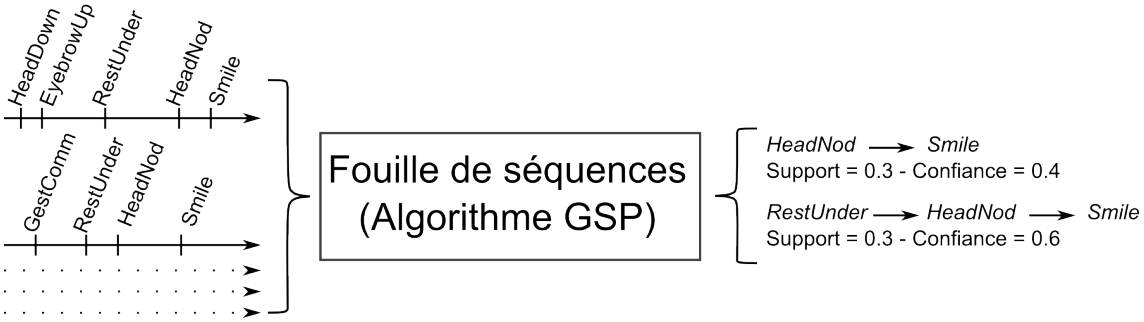


FIGURE 5.6: Application de l'algorithme de fouille de motifs séquentiels au groupe de séquences de signaux non-verbaux de chaque type de variation d'attitude. Le résultat de l'algorithme *GSP* est un ensemble de séquences fréquentes accompagnées de leur valeur de support.

L'algorithme GSP reprend le principe de l'algorithme *Apriori* de fouille de règles d'association [Agrawal & Srikant, 1994] : la première étape de l'algorithme consiste à trouver tous les évènements uniques présents au moins Sup_{min} fois dans les données. Ceux-ci peuvent être vus comme des séquences de longueur 1. Ces séquences de longueur 1 sont ensuite étendues de manière itérative, en tirant profit de la propriété que les sous-séquences d'une séquence fréquente sont elles aussi des séquences fréquentes. Nous avons utilisé la librairie de fouille de données open-source SPMF pour appliquer cet algorithme à nos données [Fournier-Viger et al., 2014].

5.2.6 Mesures de qualité

Le *support* d'une séquence représente le nombre de fois où cette séquence arrive avant un type de variation d'attitude. Cette information n'est pas suffisante pour déterminer qu'une séquence est bien caractéristique d'un type d'attitude en particulier. Par exemple, une personne prenant le tour de parole commence en général par détourner le regard, puis à rediriger son regard vers son interlocuteur au bout de quelques instants. La séquence $EyesAway \rightarrow EyesAt$ est ainsi très fréquente. Cependant, cette séquence arrive fréquemment avant tous les types de variations d'attitudes (*i.e.* sa valeur de *support* sera élevée pour tous les types de variations d'attitudes).

La dernière étape de notre méthodologie consiste ainsi à calculer d'autres mesures, appelées mesures de qualité, pour chaque séquence fréquente extraite afin de pouvoir juger de quelle variation d'attitude cette séquence est la plus caractéristique. Le lecteur peut se référer à [Tan et al., 2005] pour une description et une comparaison des nombreuses mesures permettant de caractériser des séquences fréquentes. Nous avons choisi de calculer la *confiance* et le *lift* des séquences extraites. La *confiance* nous permet de déterminer si une séquence arrive plus fréquemment avant un certain type de variation d'attitude. Le *lift* nous permet d'évaluer à quel point la relation entre une séquence et un type de variation d'attitude est inattendue. Ainsi pour les séquences de notre corpus, nous disposons des trois mesures suivantes :

- *Support*, le nombre de fois où une séquence apparait dans les données ($n \in [\![0, \infty[\![$) ;
- *Confiance*, la proportion d'occurrences qui arrivent avant un type de variation d'attitude particulier par rapport au nombre d'occurrences total de cette séquence dans les données ($x \in [0, 1]$, 1 indiquant que cette séquence arrive exclusivement avant le type de variation d'attitude considéré) ;
- *Lift*, qui indique à quel point la confiance d'une séquence est forte compte tenu des fréquences observées des évènements de la séquence, c'est à dire à quel point la confiance d'une séquence est forte par rapport à la probabilité de la co-occurrence de la séquence et de la variation d'attitude si elles étaient indépendantes ($x \in]0, \infty[$, une valeur plus haute représentant une règle plus inattendue).

Dans la section suivante, nous analysons les résultats de l'application de la méthodologie que nous venons de présenter au corpus d'entretiens d'embauche détaillé dans le chapitre 4.

5.3 APPLICATION À NOTRE CORPUS ET ANALYSE DES SÉQUENCES EXTRAITES

Pour appliquer notre méthodologie d'extraction de séquences caractéristiques de variations d'attitude à notre corpus, nous avons d'abord défini une valeur de support minimal (Sup_{min}, voir Section 5.2.5). Un compromis était nécessaire entre une valeur suffisamment grande, pour que les séquences extraites soient représentatives du comportement des recruteurs, et une valeur suffisamment faible, pour extraire des séquences plus rares mais potentiellement fortement caractéristiques d'une variation d'attitude. Une difficulté de notre problème est que nous ne disposons pas d'autant de segments pour chaque type de variation d'attitude : par exemple, nous n'observons dans notre corpus que 24 segments précédents de fortes baisses de dominance, à comparer aux 80 segments pour des faibles baisses de dominance (voir Table 5.2).

Nous avons alors décidé d'appliquer l'algorithme GSP sur les différents groupes de segments de signaux non-verbaux avec un support minimal Sup_{min} égal à 10% du nombre total de segments dans le groupe considéré (arrondi au supérieur). Ainsi, les séquences extraites pour les fortes baisses de dominance doivent être présentes au moins trois fois dans cet ensemble (24 segments, arrondi au supérieur), alors que les séquences extraites pour les faibles baisses de dominance doivent être présentes au moins huit fois (80 segments).

5.3.1 Description des résultats de l'extraction de séquences sur notre corpus

Les résultats de l'application de cette méthode à notre corpus multimodal sont présentés dans la table 5.2. Nous récupérons au total 879 séquences pour les variations de dominance et 329 pour les variations d'amicalité. En moyenne, les séquences pour les variations d'amicalité contiennent 2,91 signaux, alors que les séquences de va-

Type de variation	Centre de partition	Nombre de segments	Nombre de séquences fréquentes
Forte augmentation d'amicalité	0.34	68	86
Faible augmentation d'amicalité	0.12	66	72
Faible baisse d'amicalité	-0.11	77	104
Forte baisse d'amicalité	-0.32	36	67
Total amicalité		247	329
Forte augmentation de dominance	0.23	49	141
Faible augmentation de dominance	0.11	66	244
Faible baisse de dominance	-0.12	80	134
Forte baisse de dominance	-0.34	24	361
Total dominance		219	879

TABLE 5.2: Résultats pour chaque type de variation d'attitude. Le centre de partition correspond au résultat de l'application de l'algorithme *K-means* sur l'ensemble des variations d'attitude : le type d'une variation d'attitude d'amplitude α correspondra donc à la partition dont le centre est le plus proche de α.

riations de dominance contiennent en moyenne 3,58 signaux. On peut supposer que cette différence, à la fois dans le nombre et la longueur des séquences extraites, est à attribuer au fait que les intervalles entre deux variations de dominance sont en général plus longs (12.7 secondes) qu'entre deux variations d'amicalité (8.3 secondes : voir Table 4.6).

Séquence	Type de variation	*Support*	*Confiance*	*Lift*
BodyStraight -> ObjectManip	Forte baisse d'amicalité	13	0.31	2.09
HeadNod -> Smile	Forte augmentation d'amicalité	32	0.59	2.09
HeadNod -> RestHandsTogether -> Smile	Forte baisse de dominance	13	0.31	2.90
EyebrowsUp -> RestOverTable	Forte augmentation de dominance	21	0.33	1.54

TABLE 5.3: Exemples de séquences fréquentes obtenues par notre méthodologie de fouille de séquences de signaux non-verbaux.

Nous représentons les séquences extraites par notre méthodologie sous la forme suivante : $Signal_1 \rightarrow ... \rightarrow Signal_i \rightarrow ... \rightarrow Signal_n$, où les flèches $\rightarrow$ représentent la relation d'ordre entre les différents signaux d'une séquence, c'est à dire qu'un signal

à droite *commence* après les signaux à sa gauche. Notons que l'algorithme GSP peut extraire des séquences dont certains éléments sont des combinaisons de signaux (*ex.* $GestComm \rightarrow < Smile, HeadNod > \rightarrow RestArmsCrossed$), cependant aucune des séquences fréquentes obtenues en appliquant notre méthode sur le corpus multimodal d'entretiens d'embauche n'a présenté cette propriété. Dans la table 5.3, nous présentons des exemples de séquences extraites.

5.3.2 Analyse des séquences extraites

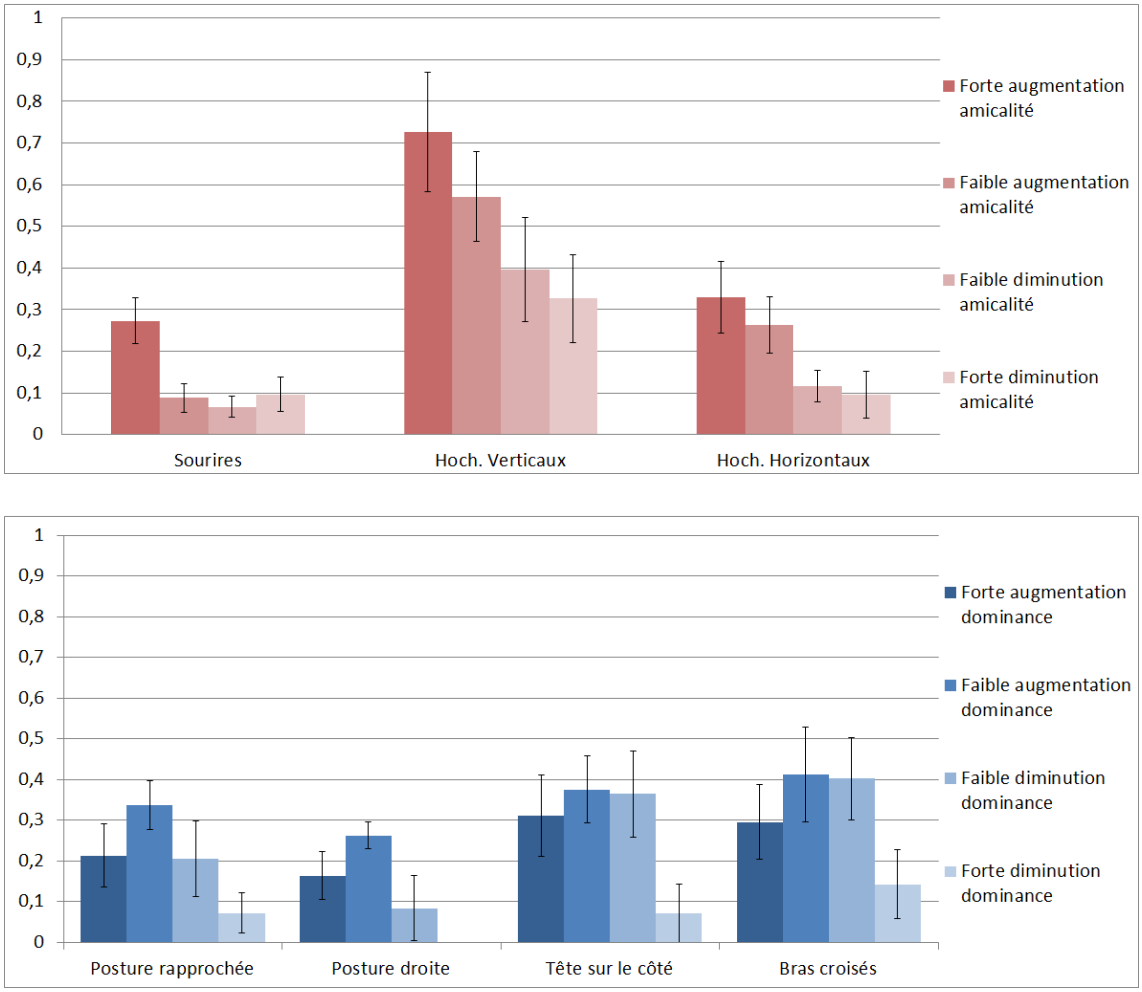


FIGURE 5.7: Comparaison des fréquences d'apparition de signaux caractéristiques d'une attitude dans les séquences fréquentes obtenues par fouille de données ($Hoch. = Hochements$). En haut : amicalité. En bas : dominance.

L'analyse des séquences extraites par des techniques de fouille de motifs séquentiels

est typiquement réalisée par un expert du domaine concerné, qui va alors explorer les séquences et décider celles qui sont dignes d'intérêt [Miller et al., 2014]. Ce travail nécessite un expert disposant de compétences préalables et demande une grande quantité de temps. Nous avons donc adopté une approche différente afin de comparer globalement les séquences extraites pour chaque type de variations d'attitude. Nous avons voulu vérifier si les signaux typiquement associés à une attitude étaient présents plus souvent dans les séquences extraites pour les types de variations correspondantes. Par exemple, le sourire étant un signe d'amicalité, il devrait être plus présent dans les séquences extraites avant des augmentations d'amicalité. Nous avons ainsi, pour chaque séquence fréquente extraite, calculé les fonctions indicatrices $\phi_{Sig}(Seq)$, c'est à dire $\phi_{Sig}(Seq) = 1$ si le signal Sig est présent dans la séquence Seq et $\phi_{Sig}(Seq) = 0$ sinon. Par exemple, dans la séquence $HeadNod \rightarrow Smile$, $\phi_{Smile}(HeadNod \rightarrow Smile) = 1$ et $\phi_{BodyLean}(HeadNod \rightarrow Smile) = 0$.

Nous avons ensuite réalisé, pour chaque type de signal, des tests de Student entre les distributions de ϕ_{Sig} obtenues pour les différents types de variations d'attitude. Nous remarquons que les sourires étaient significativement plus fréquents avant des fortes augmentations d'amicalité que dans tous les autres cas (Faibles augmentations $p = 0.005 < 0.05$, faibles baisses $p = 0.001 < 0.05$, fortes augmentations $p = 0.011 < 0.05$). Les hochements de têtes positifs (verticaux) sont significativement plus fréquents avant les fortes augmentations d'amicalité qu'avant les fortes baisses ($p = 0.026 < 0.05$). Étonnamment, la même chose est observée pour les hochements de tête négatifs (horizontaux), plus fréquents avant les fortes augmentations d'amicalité qu'avant les faibles ($p = 0.023 < 0.05$) ou fortes ($p = 0.024 < 0.05$) baisses d'amicalité.

Se pencher en avant (vers le candidat) est plus fréquent avant les faibles augmentations de dominance qu'avant les fortes baisses ($p = 0.013 < 0.05$). De manière similaire, adopter une posture droite était plus fréquent avant les faibles augmentations de dominance qu'avant les faibles baisses de dominance ($p = 0.040 < 0.05$) et les fortes augmentations ($p = 0.001 < 0.05$). Orienter sa tête vers le coté était plus fréquent avant les faibles augmentations de dominance qu'avant les fortes baisses ($p = 0.019 < 0.05$). Nous observons la même chose pour les croisements de bras ($p = 0.044 < 0.05$). Tourner la tête vers le côté semble indiquer plus rarement des fortes baisses de dominance que les autres types de variations d'attitude ($p < 0.05$ dans les trois cas).

Conclusion

Dans ce chapitre, nous avons présenté plusieurs méthodes permettant l'analyse de données séquentielles. La fouille de motifs séquentiels permet d'extraire des séquences d'évènements fréquentes dans un corpus multimodal, même si d'autres évènements indépendants sont intercalés avec les évènements de ces séquences. En cela, nous préférons cette méthode à l'analyse séquentielle par décalage et aux modèles Markoviens. Les *t-patterns* permettent aussi de modéliser ce type de séquences, cependant les relations temporelles entre les évènements de *t-patterns* seraient trop contraignantes dans un contexte de planification de comportement d'ACAs.

Nous avons donc choisi d'utiliser la fouille de motifs séquentiels pour extraire des séquences de signaux non-verbaux caractéristiques d'attitudes dans notre corpus. Nous avons proposé une méthodologie permettant de regrouper automatiquement des ensembles de séquences de signaux non-verbaux présents dans le corpus en fonction du type de variation d'attitude qu'ils précèdent. Un algorithme de fouille de motifs séquentiels est ensuite appliqué sur chacun de ces ensembles, ce qui nous permet de récupérer un ensemble de séquences fréquentes de signaux non-verbaux pour chaque type de variation d'attitude. Ces séquences fréquentes sont ensuite caractérisées par différentes mesures de qualité.

Nous avons vérifié que des signaux non-verbaux typiquement associés à l'expression d'attitudes étaient bien présents dans les séquences extraites par notre méthode. Nous remarquons notamment que les sourires et hochements de tête verticaux sont effectivement plus fréquents avant les augmentations d'amicalité.

Dans le chapitre suivant, nous proposons un modèle de planification de signaux non-verbaux pour l'expression d'attitudes utilisant les séquences fréquentes extraites par la méthodologie que nous venons de présenter dans ce chapitre.

Synthèse du chapitre

1. Différentes techniques permettent l'analyse et la modélisation de relations séquentielles entre évènements. Nous avons choisi d'utiliser la fouille de motifs séquentiels.
2. Nous avons introduit une méthodologie pour extraire automatiquement des séquences de signaux non-verbaux fréquentes avant certains types de variations d'attitudes. Cette méthodologie segmente les données par rapport aux instants de variations des attitudes, puis applique un algorithme de fouille de motifs séquentiels.
3. L'application de cette méthodologie à notre corpus nous a permis d'obtenir, pour différents types de variation d'attitude, un ensemble de séquences fréquentes accompagnées de mesures de qualité.
4. Une analyse des distributions des signaux non-verbaux présents dans les séquences fréquentes extraites confirme des résultats connus dans la littérature : par exemple, les sourires sont un signe d'amicalité.

6 Modèle computationnel de planification de séquences de signaux non-verbaux pour l'expression d'attitudes

Dans le cadre de systèmes d'entrainement aux entretiens d'embauche, il est essentiel de pouvoir confronter les utilisateurs aux différentes attitudes que pourraient adopter de réels recruteurs. Si des modèles d'expression d'attitudes par des Agents Conversationnels Animés ont été proposés (voir Section 3.2), il n'existe pas à ce jour de modèle d'expression d'attitudes sociales considérant des séquences de signaux non-verbaux. Dans ce travail de thèse, nous introduisons un modèle permettant l'expression d'attitudes sociales par un Agent Conversationnel Animé par le biais de séquences de signaux non-verbaux. Notre modèle utilise un ensemble de séquences caractéristiques de différentes attitudes extraites avec la méthodologie présentée au chapitre 5.

Nous avons conçu ce modèle pour être compatible avec l'architecture d'Agents Conversationnels Animés SAIBA, dans laquelle celui-ci remplit le rôle d'un *planificateur de comportement*. L'architecture SAIBA standardise les formats de représentation des fonctions communicatives et des signaux non-verbaux. Respecter ces standards nous permet de rendre notre modèle compatible avec d'autres composants logiciels du domaine des Agents Conversationnels Animés. Les séquences de signaux non-verbaux calculées par notre modèle pourront ainsi être utilisées pour animer les personnages virtuels de n'importe quelle plate-forme d'Agents Conversationnels Animés compatible avec SAIBA.

Dans ce chapitre, nous commençons par présenter l'architecture d'Agents Conversationnels Animés SAIBA et ses langages de représentation des fonctions communicatives et des comportements. La deuxième section détaille le modèle de planification que nous proposons et les différentes étapes pour la génération de séquences de signaux non-verbaux exprimant une attitude. Nous terminons le chapitre en présentant une évaluation réalisée afin de vérifier que les séquences de signaux calculées

par notre modèle parviennent à exprimer les attitudes voulues.

6.1 PRÉSENTATION DE SAIBA

SAIBA (Situation, Agent, Intention, Behavior, Animation) est une architecture de référence dans le domaine des Agents Conversationnels Animés [Vilhjálmsson et al., 2007], issue d'une série de collaborations internationales (*Representations for Multimodal Generation Workshop*, 2005, *HUMAINE Joint Workshop on Representations for Multimodal Behavior*, 2006) [Kopp et al., 2006]. Cette architecture sépare les fonctions de planification d'intentions communicatives (*Intent Planning*), de planification de comportement (*Behavior Planning*), et de réalisation de ces comportements (*Behavior Realization*). Des langages de représentation des fonctions communicatives (FML : *Function Markup Language* [Heylen et al., 2008]) et du comportement (BML : *Behavior Markup Language* [Vilhjálmsson et al., 2007]) ont été proposés afin de standardiser les entrées et sorties des composants logiciels remplissant une des trois fonctions de SAIBA. L'architecture SAIBA est représentée sur la figure 6.1.

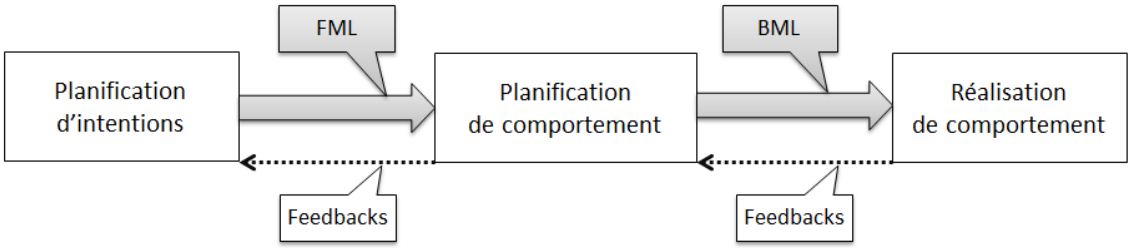


FIGURE 6.1: Architecture SAIBA.

Le langage FML est utilisé pour représenter des fonctions communicatives [Heylen et al., 2008], indépendamment des signaux qui vont ensuite être utilisés par un ACA pour les réaliser. Le langage FML est à ce jour encore en cours de standardisation [Cafaro et al., 2014]. Ainsi, les catégories de fonctions communicatives incluses dans un fichier FML varient selon le domaine d'application ou le groupe de recherche. Un exemple de fichier FML pour la phrase « Quelles sont les choses qui vous passionnent au travail ? » est présenté en figure 6.2. Cette phrase est annotée avec l'intention communicative de poser une question (balise `<performative type="ask">`). La balise `<pitchaccent>` représente l'accentuation du mot *passionnent* et la balise

`<boundary>` marque un changement d'intonation indiquant la fin de la phrase (marqueurs *méta-linguistiques* de [Poggi, 2003]). Des marqueurs temporels (`timemarkers`) sont utilisés pour synchroniser les intentions communicatives avec le discours.

```
<?xml version="1.0"?>
<fml-apml>
  <bml id="ask-motivations-F" agent="Greta">
    <speech id="s1" language="french" start="0.0" voice="cereproc">
      <timemarker id="tm1"/>
        Quelles sont les choses qui vous
      <timemarker id="tm2"/>
        passionnent
      <timemarker id="tm3"/>
        au travail ?
      <timemarker id="tm4"/>
        <pitchaccent start="tm2" end="tm3" id="pa1" importance="1"
            level="medium" type="Hstar"/>
        <boundary start="tm4" end="tm4+0.5" id="b1" type="HH"/>
    </speech>
  </bml>
  <fml>
    <performative id="p1" type="ask" end="tm4" start="tm1"
        importance="1"/>
  </fml>
</fml-apml>
```

FIGURE 6.2: Exemple de fichier FML.

Les signaux non-verbaux sont eux exprimés dans le format BML [Vilhjálmsson et al., 2007]. A l'inverse du langage FML, celui-ci a atteint un premier niveau de standardisation[1]. Un exemple de fichier BML est présenté dans la figure 6.3. Dans le format BML, les signaux non-verbaux sont représentés par des balises correspondant à la modalité d'expression du signal (*ex.* `<gaze>` pour le regard). Des attributs sont utilisés pour spécifier le type du signal (`lexeme`) et d'autres paramètres (*ex.* le début et la fin du signal : `start` et `end`).

Dans ce travail de thèse, nous introduisons un modèle de planification de comportement pour l'expression d'attitudes. Celui-ci reçoit ainsi en entrée un fichier FML décrivant des fonctions communicatives, et fournit en sortie un fichier BML décrivant les signaux non-verbaux qui réalisent ces fonctions communicatives. De plus,

[1] `http://www.mindmakers.org/projects/bml-1-0/wiki`

```
<?xml version="1.0"?>
<bml xmlns="http://www.mindmakers.org/projects/BML" character="Greta"
    id="bml1">
    <speech id="s1" language="french" start="0.0" voice="cereproc">
        <timemarker id="tm1"/>
            Quelles sont les choses qui vous
         <timemarker id="tm2"/>
            passionnent
         <timemarker id="tm3"/>
            au travail ?
         <timemarker id="tm4"/>
        <pitchaccent start="tm2" end="tm3" id="pa1" importance="1"
            level="medium" type="Hstar"/>
        <boundary start="tm4" end="tm4+0.5" id="b1" type="HH"/>
    </speech>
    <head start="0.984" end="1.594" id="h1" lexeme="Down_Aside_Right" />
    <gesture ready="0.0" relax="2.260" id="g1" lexeme="ask_Ges_B"/>
    <face start="1.594" end="2.260" id="g1" amount="0.821" lexeme="Smile"/>
    <gaze start="0.0" id="r1" target="User"/>
    <torso start="0.00" end="4.772" id="t1" lexeme="BodyLean" >
</bml>
```

FIGURE 6.3: Exemple de fichier BML.

une autre entrée de modèle est une variation d'attitude que l'ACA doit exprimer. L'objectif de notre modèle est de générer une séquence de signaux non-verbaux exprimant à la fois les intentions communicatives du fichier FML et l'attitude désirée. Dans la prochaine section, nous présentons notre modèle et les différentes étapes de la génération de séquences de signaux non-verbaux.

6.2 MODÈLE DE PLANIFICATION DE SÉQUENCES DE SIGNAUX NON-VERBAUX

Le modèle de planification de séquences de signaux non-verbaux pour l'expression d'attitudes que nous proposons comporte trois étapes, qui sont détaillées dans les sections suivantes. Premièrement, pour chaque intention communicative contenue dans le fichier FML donné en entrée, nous faisons l'inventaire des signaux pouvant être utilisés pour l'exprimer. En choisissant un de ces signaux pour chacune des intentions du fichier FML, nous construisons une séquence de signaux exprimant ces

intentions. Nous appelons ces premières séquences des *séquences minimales* (Section 6.2.1).

Ensuite, pour chaque *séquence minimale* obtenue dans l'étape précédente, notre modèle repère ensuite tous les intervalles de temps où d'autres signaux peuvent être insérés. En insérant des signaux additionnels dans ces intervalles, nous construisons alors de nouvelles séquences plus longues, appelées *séquences candidates*. Ces signaux additionnels vont permettre à l'agent d'exprimer une attitude (Section 6.2.2).

Enfin, la troisième étape consiste à sélectionner la meilleure séquence (*i.e.* la séquence présentant la plus grande probabilité d'exprimer la variation d'attitude voulue) parmi les *séquences candidates* par le biais d'une méthode de classification utilisant les séquences extraites par la méthodologie présentée au chapitre 5.

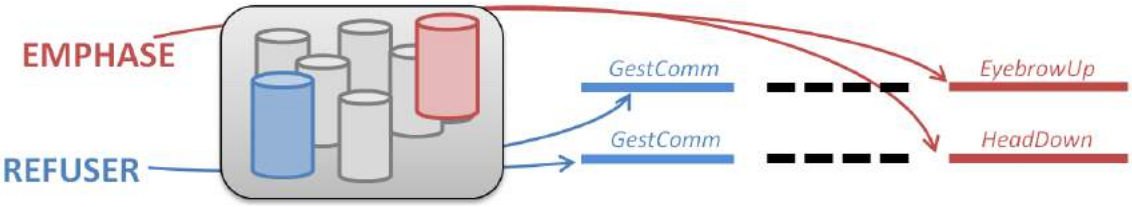
Les prochaines sections détaillent ces trois étapes. Au cours de nos explications, nous prendrons l'exemple de la planification d'une séquence de signaux non-verbaux pour la phrase « Vous n'avez pas les compétences requises pour le poste auquel vous avez postulé. », en supposant que la variation d'attitude à exprimer est une faible baisse de dominance. Cette phase est exprimée dans un fichier FML contenant plusieurs fonctions communicatives : pour notre exemple, l'ACA doit exprimer qu'il refuse le poste au candidat, et mettre l'emphase sur une partie de la phrase. Le fichier FML est représenté graphiquement dans la partie (a) de la figure 6.4.

6.2.1 Construction de séquences minimales

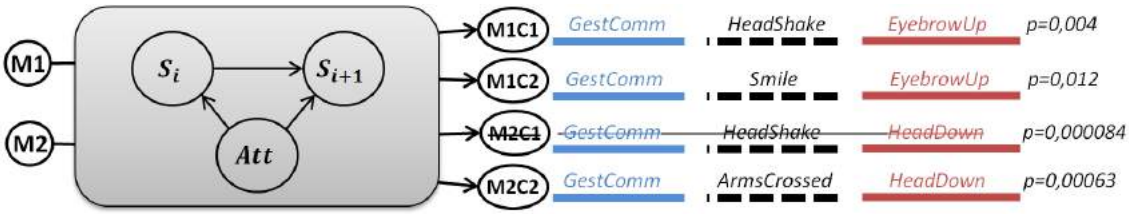
Lors de conversations, les intentions communicatives des participants sont exprimées au travers du discours mais aussi du comportement non-verbal [Poggi, 2003]. Par exemple, pour mettre l'accent sur une partie du discours, il est courant de hausser les sourcils ou de faire un léger mouvement de tête vers le bas [Condon & Osgton, 1971]. Le langage FML [Heylen et al., 2008] représente ces intentions communicatives.

La première étape de notre modèle consiste à récupérer tous les signaux non-verbaux pouvant exprimer les intentions contenues dans le message FML reçu en entrée. Pour cela, nous utilisons la méthode décrite par Mancini [Mancini & Pelachaud, 2008]. Celle-ci consiste à définir un lexique dans lequel chaque intention communicative est caractérisée par un *Behavior Set*. Ceux-ci spécifient les différents signaux non-verbaux qui peuvent être utilisés pour exprimer une intention, ainsi que

(a) Exemple de message FML en entrée de notre modèle.



(b) Génération de séquences minimales utilisant les *Behavior Set* de chaque intention communicative du message donné en entrée (*i.e.* *REFUSER* et *EMPHASE*). Dans notre exemple, nous obtenons deux séquences minimales, **M1** et **M2**.



(c) Les séquences minimales (**M1** et **M2** sont enrichies, créant un ensemble de séquences candidates (**M1C1**, **M1C2**, **M2C1** et **M2C2**). Les probabilités de ces séquences calculées avec les réseaux Bayésiens permettent de retirer les séquences trop improbables (*ex.* ici, **M2C1**).

(d) Sélection de la meilleure séquence parmi les séquences candidates, *i.e.* la séquence s avec le score $Sc(s)$ le plus élevé.

FIGURE 6.4: Représentation graphique du modèle de génération de séquences.

des contraintes éventuelles sur ces signaux. Dans le cadre de ce travail, les intentions communicatives que nous considérons sont inspirées de la taxonomie de Poggi (voir Section 2.1.3) [Poggi, 2003]. Nous considérons par exemple l'expression d'intentions performatives (*ex.* un ACA peut vouloir *informer*, *suggérer*, *demander*, *saluer*, *etc.*), les fonctions *méta-linguistiques* et *méta-discursives* (*ex.* mettre l'*accent* sur un mot, indiquer qu'une phrase est terminée) et l'expression d'informations spatio-temporelles (*ex.* emplacement d'un objet, qui pourra être réalisée par un geste déictique). Dans le cadre de notre groupe de recherche, différents projets et travaux réalisés au cours des années précédents ont permis de rassembler un lexique contenant un *Behavior Set* pour de nombreuses intentions communicatives.

La première étape de notre modèle consiste à construire des *séquences minimales*, c'est à dire des séquences qui contiennent un signal non-verbal pour chaque intention communicative à exprimer. Nous choisissons de ne considérer qu'un seul signal non-verbal par intention communicative car aucune séquence fréquente extraite de notre corpus multimodal ne contenait de combinaison de signaux simultanés (voir Section 5.3.1). Ce choix constitue une limite de notre méthode, que nous détaillons dans la section 9.2.3. Pour construire une *séquence minimale*, il suffit donc de sélectionner, pour chaque intention communicative du fichier FML à exprimer, un des signaux du *Behavior Set* correspondant. Cette première étape permet de s'assurer que toutes les intentions communicatives du fichier FML seront exprimées dans les séquences de signaux non-verbaux que nous allons générer. Cela permet, en particulier, d'obtenir les formes de gestes qui vont correspondre aux intentions du discours. Nous obtenons alors l'ensemble des *séquences minimales* en construisant toutes les séquences possibles comprenant un signal pour chaque intention communicative du message FML initial.

Dans le cas de l'exemple que nous avons proposé, le lexique devra contenir une définition des signaux non-verbaux pouvant exprimer l'intention performative de *refuser* quelque chose (*i.e.* ici, refuser le poste au candidat), et une définition des signaux non-verbaux pouvant être utilisés pour placer l'*emphase* sur une partie du discours. Nous prenons en exemple le lexique présenté dans la figure 6.5. Dans cet exemple, la fonction communicative *refuser* peut être réalisée par un geste (`Wipe`). Pour placer l'emphase sur un mot, deux signaux peuvent être utilisés : un mouvement de tête (`Down_Aside`) ou un haussement de sourcils (`Raise_Eyebrows`). Nous pouvons donc construire deux séquences minimales, représentées dans la partie (b) de la figure 6.4 :

un geste puis un haussement de sourcis (**M1**), ou bien un geste puis un mouvement de tête vers le bas (**M2**).

```
<?xml version="1.0"?>
<behaviorsets>
    ...
    <behaviorset name="performative-refuser">
        <signal id="1" name="Wipe" modality="gesture"/>
    </behaviorset>
    ...
    <behaviorset name="emphase">
        <signal id="1" name="Down_Aside" modality="head"/>
        <signal id="2" name="Raise_Eyebrows" modality="face"/>
    </behaviorset>
    ...
</behaviorsets>
```

FIGURE 6.5: Exemple de lexique indiquant les signaux pouvant être utilisés pour réaliser une fonction communicative.

La prochaine étape consiste à enrichir ces séquences avec des signaux additionnels qui vont permettre d'exprimer l'attitude désirée.

6.2.2 Génération de séquences candidates

Pour chaque *séquence minimale* obtenue à l'étape précédente, nous commençons par repérer tous les intervalles temporels où il est possible d'insérer d'autres signaux. Par exemple, s'il y a assez de temps entre deux signaux, nous pourrions insérer un hochement de tête, un sourire, ou encore un changement de posture.

Afin de représenter la relation entre signaux adjacents et utiliser cette connaissance pour l'insertion de nouveaux signaux, nous avons construit deux réseaux Bayésiens, un pour les variations de dominance et l'autre pour les variations d'amicalité. Les noeuds de ces réseaux représentent les signaux non-verbaux successifs et les attitudes sociales (Figure 6.6). Les arêtes de ces réseaux définissent les relations entre deux variables. Ces réseaux Bayésiens nous permettent de représenter la relation causale et non-déterministe des attitudes sur les signaux (*ex.* il pourrait y avoir plus de sourires pour les augmentations d'amicalité, ou plus de bras croisés pour les baisses

d'amicalité) et des séquences de signaux (*ex.* des changements de position de repos des mains arriveront souvent après des gestes).

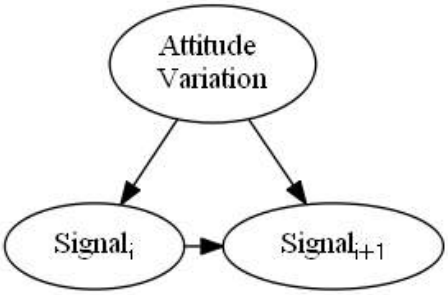


FIGURE 6.6: Représentation du réseau Bayésien que nous utilisons pour générer des séquences candidates.

Nous notons que $P(S_{i+1}|S_i, S_{i-1}, ..., S_1, A) = P(S_{i+1}|S_i, A)$, où $\{S_1, ..., S_i, S_n\}$ représentent les signaux de la séquence, i l'indice du i-ème signal, et A la variation d'attitude voulue. On peut noter que certains signaux ne peuvent être adjacents (les signaux de changements de posture, de position de tête ou de position de repos des mains ne peuvent pas arriver après un changement identique, *ex.* $HeadAt \rightarrow HeadAt$ ou $BodyStraight \rightarrow BodyStraight$), et nous nous sommes assurés que ces chemins n'existent pas dans les réseaux Bayésiens que nous avons construits.

Pour apprendre les paramètres de ces réseaux bayésiens, nous avons utilisé le logiciel open-source d'apprentissage automatique Weka [Hall et al., 2009], en utilisant notre corpus multimodal comme base de données d'apprentissage. Chaque paire de signaux non-verbaux adjacents observée avant un type de variation d'attitude est ainsi un exemple utilisé pour l'apprentissage des probabilités de nos réseaux Bayésiens.

Les réseaux Bayésiens obtenus sont utilisés pour insérer de nouveaux signaux aux *séquences minimales* en fonction de l'attitude donnée en entrée. Étant donné une *séquence minimale* en entrée, la taille maximale des séquences générées est égale au nombre de signaux présents dans la *séquence minimale* plus le nombre d'intervalles vides où de nouveaux signaux peuvent être insérés. Afin de réduire la complexité combinatoire de cette étape et d'élaguer les séquences qui sont trop improbables (c'est à dire contenant des paires de signaux adjacents très peu observées dans les données avant un type d'attitude), nous calculons la probabilité de chaque séquence générée après l'insertion d'un nouveau signal, et conservons uniquement les séquences dont la probabilité obtenue dépasse un certain seuil λ. Afin que ce seuil

dépende de la probabilité initiale de la *séquence minimale*, nous avons choisi que λ soit égal à $P(SéquenceMinimale) * \alpha$ où α est un coefficient qui permet d'ajuster la quantité de séquences générées. Le choix de la valeur de α dépend du nombre de signaux à considérer et de la taille maximale des séquences à générer. Dans notre cas, nous avons trouvé après des essais préliminaires que $\alpha = 0.005$ était un compromis adéquat entre la quantité de séquences générées et le temps de calcul.

Dans l'exemple que nous avons proposé, un seul intervalle est disponible pour insérer de nouveaux signaux non-verbaux (dans la figure 6.4(a), le segment entre les intentions *REFUSER* et *EMPHASE*). Nous créons alors de nouvelles séquences à partir des *séquences minimales* en y ajoutant d'autres signaux non-verbaux dans cet intervalle. Par exemple, dans la figure 6.4(c), nous montrons la génération des *séquences candidates* $GestComm \rightarrow HeadShake \rightarrow EyebrowUp$ (**M1C1**) et $GestComm \rightarrow Smile \rightarrow EyebrowUp$ (**M1C2**) à partir de la *séquence minimale* **M1**, et des *séquences candidates* $GestComm \rightarrow HeadShake \rightarrow HeadDown$ (**M2C1**) et $GestComm \rightarrow ArmsCrossed \rightarrow HeadDown$ (**M2C2**) à partir de la *séquence minimale* **M2**. En pratique, nous générons une nouvelle séquence en essayant d'insérer tous les types de signaux non-verbaux annotés dans notre corpus multimodal (voir Table 4.4). Toutes ces *séquences candidates* sont évaluées en utilisant le réseau Bayésien correspondant à l'attitude à exprimer. Dans notre exemple, comme nous voulons que l'ACA exprime une baisse de dominance, nous utilisons le réseau Bayésien construit pour les variations de dominance pour estimer les probabilités de ces nouvelles séquences. Les séquences dont la probabilité est trop faible, dans notre exemple la séquence **M2C1**, sont abandonnées. Dans notre exemple, la séquence **M2C1** est retirée, car sa probabilité est très faible (on peut imaginer, par exemple, que la probabilité des deux mouvements de tête adjacents $HeadShake$ puis $HeadDown$ soit très faible et que cela soit la raison de la faible probabilité de la séquence **M2C1**).

Les séquences enrichies de nouveaux signaux après cette étape sont appelées les *séquences candidates*. L'étape suivante consiste à choisir la meilleure séquence dans l'ensemble des *séquences candidates*, c'est à dire la séquence la plus à même d'exprimer la variation d'attitude voulue.

6.2.3 Sélection de la séquence finale

Une fois que l'ensemble de *séquences candidates* a été généré, notre modèle détermine quelle séquence de cet ensemble a le plus de chances d'exprimer l'attitude voulue.

Pour identifier cette séquence, nous nous inspirons d'une méthode de classification de texte par vote à la majorité (*majority voting*) à partir de séquences fréquentes proposée par Jaillet *et al.* [Jaillet et al., 2006]. Pour chaque séquence candidate s, nous commençons par extraire les k sous-séquences de signaux non-verbaux sub_i contenues dans s qui possèdent les plus hautes mesures de *confiance* parmi toutes les séquences extraites par fouille de données, indépendamment du type de variation d'attitude pour lequel ces sous-séquences ont été extraites. Ces sous-séquences vont ensuite contribuer à hauteur d'une voix en faveur du type de variation d'attitude pour lequel elles ont été extraites. On compte alors le nombre de voix par type de variation d'attitude, et la séquence s est classifiée comme exprimant le type de variation ayant récolté le plus de voix. Si on arrive à une égalité de voix entre plusieurs types de variations d'attitude, nous choisissons celle dont la moyenne de la confiance de ses sous-séquences est la plus forte. Les séquences candidates qui sont classifiées comme exprimant un autre type de variation d'attitude que celle voulue en entrée sont retirées.

Voici un exemple de classification de séquence de signaux non-verbaux par cette méthode de vote à la majorité. Supposons qu'on cherche à classifier la séquence **M1C2** ($GestComm \rightarrow Smile \rightarrow EyebrowUp$) en terme de variation de dominance. On parcourt l'ensemble des séquences fréquentes extraites pour les différents types de variations de dominance qui sont des sous-séquences de **M1C2**, et on garde les k sous-séquences avec la plus forte valeur de confiance. Par exemple, en utilisant $k = 3$, on peut obtenir les sous-séquences suivantes :

$GestComm \rightarrow EyebrowUp$	Forte augmentation	$Confiance = 0.47$
$Smile$	Faible baisse	$Confiance = 0.43$
$Smile \rightarrow EyebrowUp$	Faible baisse	$Confiance = 0.38$

Chacune de ces sous-séquences contribue alors d'une voix envers son type de variation d'attitude à la classification de la séquence **M1C2**. On obtient ici deux voix pour une faible baisse de dominance, et une voix pour une forte augmentation ; la sé-

quence **M1C2** est alors classifiée comme exprimant une faible baisse de dominance. C'est la variation d'attitude voulue : la séquence **M1C2** est donc conservée.

Nous calculons ensuite un score $Sc(s)$ pour chaque séquence s conservée après l'étape précédente. En posant $FSeq$ l'ensemble des séquences fréquentes extraites par fouille de données, et $\lambda_s = 1$ si $s \in FSeq$, $\lambda_s = 0$ sinon , on définit $Sc(s)$ ainsi :

$$Sc(s) = \begin{cases} Support(s) * Confiance(s), & \text{si } s \in FSeq \\ \sum_{i=1}^{k} \lambda_{sub_i} * Sc(sub_i), & \text{sinon} \end{cases}$$

Finalement, on choisit la séquence s présentant le plus haut score $Sc(s)$ parmi toutes les séquences restantes. La dernière étape consiste à exprimer cette séquence dans le format BML [Vilhjálmsson et al., 2007]. Cette séquence peut alors être interprétée par toute plate-forme d'Agent Conversationnel Animé compatible avec SAIBA afin d'animer un personnage virtuel.

Dans la prochaine section, nous présentons une première évaluation de notre modèle de planification de séquences de signaux non-verbaux pour l'expression d'attitudes.

6.3 ÉVALUATION

Afin d'évaluer ce modèle de génération de séquences, nous avons réalisé une étude destinée à vérifier que les séquences de signaux non-verbaux générées pour une certaine variation d'attitude en entrée sont finalement bien perçues comme exprimant ce type de variation d'attitude (*ex.* augmentation de dominance), et avec la bonne intensité (*ex.* forte augmentation). Dans les sections suivantes, nous commençons par décrire l'étude et nous rapportons ensuite ses résultats.

6.3.1 Description de l'étude

L'étude a été réalisée en ligne. La plateforme de l'étude a été développée avec la technologie Adobe Flash. Chaque participant à l'étude s'est vu demander de comparer huit paires de vidéos d'un personnage virtuel dans le rôle d'un recruteur, dans

lesquelles le recruteur pose une question tout en affichant des signaux non-verbaux (voir Figure 6.7).

Pour chaque paire de vidéos, le recruteur virtuel posait une question différente (*ex.* « *In your previous professional experiences, did you ever have to deal with difficult situations* ? »). Les huit questions étaient toujours présentées dans le même ordre. Les intentions communicatives contenues dans les questions ont été restreintes aux actes de dialogues performatifs suivants : *poser* une question, *informer*, *proposer* un sujet de conversation. De plus, des intentions de communication de relations spatio-temporelles étaient aussi incluses : *ici*, *maintenant*, *dans le passé*. Enfin, des marqueurs prosodiques d'emphase et de pauses ont été annotés dans ces questions.

Pour chaque paire de vidéos, le discours du recruteur était identique, et produit avec le moteur de synthèse vocale Cereproc [Aylett & Pidcock, 2007]. Le comportement non-verbal du recruteur, en revanche, était différent entre les deux vidéos. Ainsi, nous avons considéré que les réponses des participants n'étaient pas affectées par le contenu verbal des questions posées.

Chacune de ces huit paires de vidéos correspondait à une des conditions testées, c'est à dire à une des huit variations d'attitudes considérées : forte augmentation de dominance, forte baisse de dominance, forte augmentation d'amicalité, forte baisse d'amicalité, faible augmentation de dominance, faible baisse de dominance, faible augmentation d'amicalité, faible baisse d'amicalité. Sur la vidéo de droite (voir Figure 6.7), le comportement du recruteur virtuel était déterminé par notre modèle, en utilisant en entrée la phrase considérée (exprimée en FML) pour cette paire de vidéos, et la variation d'attitude de la condition testée. Sur la vidéo de gauche, le comportement du recruteur virtuel était généré avec un comportement neutre : pour cela, nous avons utilisé un planificateur de comportement existant [Mancini & Pelachaud, 2008]. Nous avons fait l'hypothèse que les vidéos générées par ce modèle, dans lequel l'expression d'attitudes n'est pas considérée, seraient considérées neutres en termes d'attitudes exprimées. En tout, ce sont ainsi 64 séquences distinctes qui ont été évaluées avec notre étude (huit questions posées par le recruteur $*$ huit attitudes), et 72 vidéos ont été générées pour réaliser l'étude (64 + 8 neutres).

Pour chaque paire de vidéos, les participants devaient répondre aux deux questions suivantes : *Q1 : « Compared to the Reference Video (left), the character on the Comparison video (right) is : »* (« Par rapport à la vidéo de référence (gauche), le

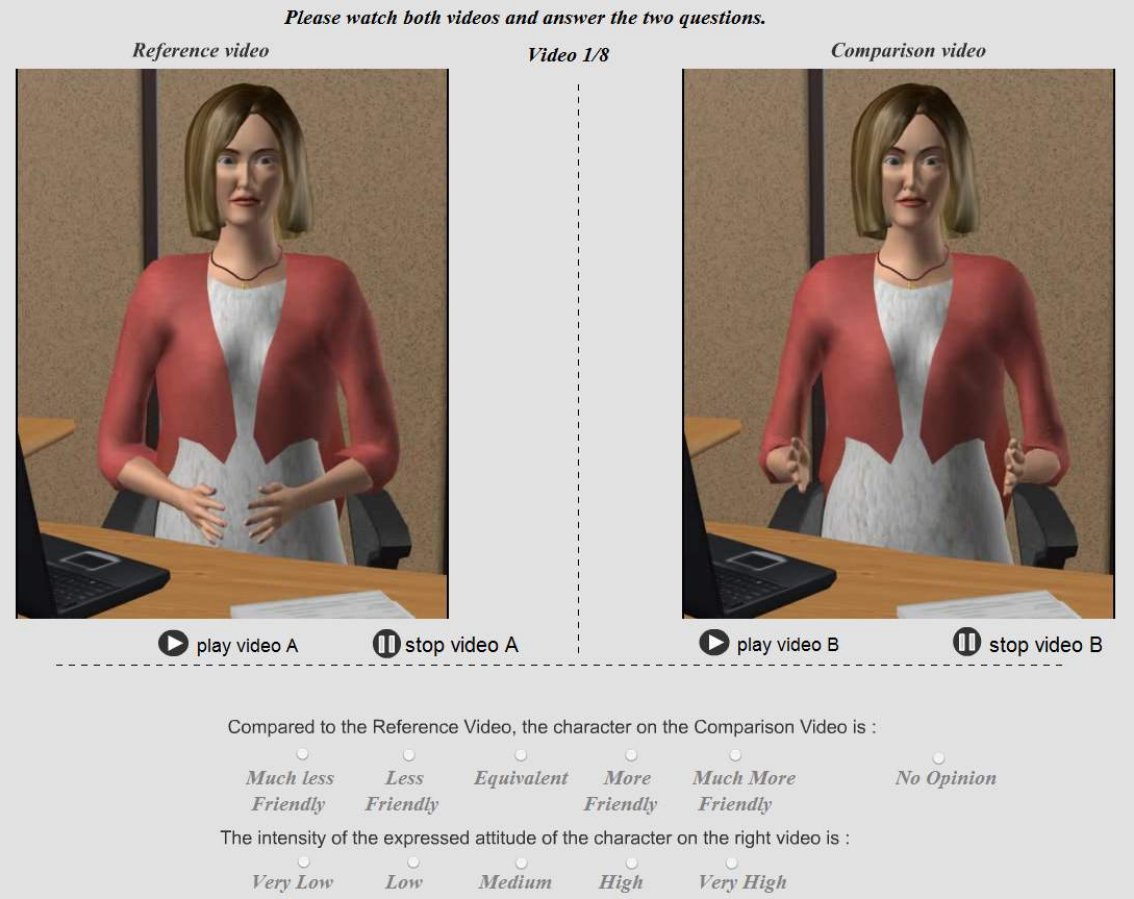


FIGURE 6.7: L'écran principal de l'étude en ligne.

personnage de la vidéo de comparaison (droite) est : »), avec les valeurs possibles suivantes : « Much less dominant », « Less dominant », « Equivalent », « More dominant », « Much more dominant », « Undecided » (« Beaucoup moins dominant », « Moins dominant », « Equivalent », « Plus dominant », « Beaucoup plus dominant » ou « Ne se prononce pas », ou les réponses équivalentes dans le cas de l'amicalité). Les participants n'ayant pas choisi « Undecided », se voyaient alors poser une deuxième question : *Q2 : « The intensity of the expressed attitude on the Comparison video (right) is : »* (« L'intensité de l'attitude exprimée dans la vidéo de comparaison (droite) est : »), avec les valeurs possibles suivantes : « Very low », « Low », « Medium », « High », « Very high » (« Très basse », « Basse », « Moyenne », « Haute », « Très haute »).

6.3.2 Résultats

Quatre-vingt-un participants ont pris part à l'étude (43 femmes, 38 hommes). Les participants étaient principalement de nationalité française (88%), et l'âge moyen de la population était de 32.4 ans (σ : 12.8).

Dans la table 6.1, nous rapportons les pourcentages de chaque type de réponse par rapport aux types de variations d'attitudes pour *Q1*. Les valeurs moyennes sont considérées en partant du principe que les réponses sont sur une échelle ordinale (Beaucoup moins amical = 1, beaucoup plus amical = 5, *etc.*). Pour juger si nos résultats étaient significatifs, nous avons réalisé des tests T de Student pour comparer si la moyenne des réponses était significativement différente de la moyenne de l'échelle ($\mu = 3$). Si une différence significative est détectée, cela indique que l'attitude exprimée dans la vidéo affichant une séquence de signaux non-verbaux calculée par notre modèle était jugée significativement différente de l'attitude exprimée par le modèle de planification de comportement que nous considérons neutre.

Les résultats sont significatifs et confirment une bonne reconnaissance pour les augmentations de dominance ($\mu = 3.55, \sigma = 0.94, t(126) = 7.09, p = 0.000$), augmentations d'amicalité ($\mu = 3.21, \sigma = 1.02, t(135) = 2.43, p = 0.016$) et diminutions d'amicalité ($\mu = 2.75, \sigma = 0.99, t(131) = 2.99, p = 0.003$). En revanche, le test de Student ne montre pas de différence significative pour les diminutions de dominance, mais une tendance semble indiquer que ces attitudes ne sont pas reconnues ($\mu = 3.09, \sigma = 0.95, t(129) = 1.16, p = 0.25$).

Pour *Q2*, nous avons réalisé des tests T de Student entre les paires de conditions du même type (*ex.* augmentation ou baisse de dominance ou d'amicalité) mais d'intensité différente (*ex.* faible ou forte). La seule différence significative a été trouvée entre les intensités perçues de fortes baisses et de faibles baisses d'amicalité. Cependant, cette différence significative révèle un résultat négatif : les séquences générées pour une plus forte baisse d'amicalité ($\mu = 2.97$) sont notées moins intenses que les séquences générées pour les faibles baisses d'amicalité ($\mu = 3.31$, $p = 0.016 < 0.05$). Les différences entre les intensités des augmentations d'amicalité ($p = 0.62$), des baisses de dominance ($p = 0.48$) et des augmentations de dominance ($p = 0.73$) ne sont pas significatives.

Dans la section suivante, nous analysons les résultats de l'étude.

	Diminution d'amicalité	Augmentation d'amicalité	Diminution de dominance	Augmentation de dominance
Beaucoup moins (1)	3.73%	2.94%	2.26%	0.78%
Moins (2)	**45.5**%	22.8%	24.1%	14.7%
Équivalent (3)	24.6%	**33.8**%	**39.1**%	20.9%
Plus (4)	20.9%	31.6%	27.1%	**52.7**%
Beaucoup plus (5)	3.73%	8.8%	5.26%	9.30%
Ne se prononce pas	1.04%	0%	2.26%	1.55%
Moyenne	2.75	3.21	3.02	3.50
Valeur t	$t(131) = 2.99$	$t(135) = 2.43$	$t(129) = 1.16$	$t(126) = 7.09$
Valeur p	0.003	0.02	0.25	0.000

TABLE 6.1: Répartition des réponses à $Q1$ pour les quatre conditions, valeurs moyennes des réponses et statistiques des tests de Student.

6.4 DISCUSSION

Le but de l'étude présentée dans la section précédente était de vérifier que notre modèle est en mesure de générer des séquences de signaux non-verbaux qui expriment une variation d'attitude donnée en entrée, et que la variation d'attitude perçue l'est avec la bonne intensité.

Les résultats de *Q1* valident partiellement notre modèle. En effet, ceux-ci indiquent que les séquences générées pour des augmentations de dominance, des augmentations et des baisses d'amicalité sont perçues correctement. Toutefois, les séquences générées pour les baisses de dominance sont perçues comme équivalentes aux vidéos de référence.

Pour *Q2*, la seule différence significative a été perçue entre les intensités de fortes et faibles baisses d'amicalité, cependant les vidéos générées pour de plus faibles variations ont été perçues comme plus intenses que celles générées pour de plus fortes variations. Il semble donc que notre modèle n'est pas en mesure d'exprimer des variations d'attitude de différentes intensités.

L'analyse de ces résultats amène plusieurs réflexions. Un facteur ayant pu influencer les résultats de *Q1* est que l'acte même de prendre la parole lors d'un entretien d'embauche peut être vu comme une prise de contrôle de l'entretien, et ainsi comme une forme d'affirmation de sa dominance sur l'interlocuteur. Le rôle de recruteur, lui aussi, est porteur de dominance car il implique que le personnage virtuel a le

pouvoir d'accorder un bénéfice à son interlocuteur (*i.e.* un emploi). On peut ainsi se demander s'il est possible pour un personnage virtuel d'exprimer une baisse de dominance tout en parlant.

Si le choix des signaux non-verbaux multimodaux était différent entre les vidéos de référence et les vidéos générées avec notre modèle, l'expressivité du comportement du recruteur (*ex.* amplitude des gestes et des mouvements de tête, intensité et durée des gestes et des expressions faciales) était la même dans les deux cas. Or ces paramètres peuvent influencer l'attitude (*ex.* des gestes larges indiquent de la dominance). De plus, il est possible que le terme « d'intensité » de l'attitude exprimée par le recruteur ait pu être mal interprété. En particulier, ce terme a pu être confondu avec l'intensité gestuelle, ce qui aura pu être déterminant pour *Q2*.

Une limite de notre protocole d'évaluation est que l'agent n'est évalué que lorsqu'il prend la parole. Or, le comportement d'écoute produit par un ACA lorsqu'une personne lui adresse la parole contribue certainement à l'expression d'attitudes. Par exemple, nous savons que mimer les comportements de son interlocuteur, comme adopter la même posture que lui [LaFrance, 1982], est un signe d'amicalité. Cependant, notre protocole d'évaluation ne permettait pas d'étudier cet effet. Une autre limite de notre protocole d'évaluation, par le biais de vidéos, est que les participants ne sont pas directement impliqués dans l'entretien d'embauche. Ceux-ci n'ont pas à répondre aux questions du recruteur virtuel. La perception des attitudes exprimées par l'ACA pourrait être différente lors d'une interaction directe, plutôt qu'en regardant des vidéos présentées séparément.

Conclusion

Dans ce chapitre, nous avons introduit un modèle de planification de séquences de signaux non-verbaux pour l'expression d'attitudes. Notre modèle s'intègre dans l'architecture d'Agents Conversationnels Animés SAIBA où il remplit la fonction de planification de comportement. Il reçoit donc en entrée un fichier FML contenant des intentions communicatives et fournit en sortie un fichier BML décrivant les signaux non-verbaux à afficher par un ACA.

Notre modèle de génération de séquences de signaux non-verbaux comporte trois étapes. La première étape consiste à construire des séquences des signaux non-

verbaux pour réaliser les intentions communicatives contenues dans le fichier FML. Ensuite, des signaux non-verbaux additionnels sont insérés dans les séquences obtenues dans l'étape précédente. La dernière étape consiste à sélectionner la séquence la plus à même d'exprimer l'attitude voulue.

Nous avons réalisé une étude pour évaluer l'expression d'attitudes par notre modèle. Des participants évaluaient une vidéo d'un ACA produisant une séquence de signaux non-verbaux générée par le biais de notre modèle par rapport à une vidéo générée avec un modèle de planification de comportement existant, qui ne considérait pas l'expression d'attitudes. Cette étude a permis de valider que les variations d'amicalité, d'inamicalité et de dominance exprimées par notre modèle sont correctement reconnues par les participants. Cependant, les baisses de dominance ne sont pas reconnues, et les intensités de variation d'attitude ne sont pas non plus reconnues.

Dans le prochain chapitre, nous présentons notre implémentation d'un recruteur virtuel autonome utilisant notre modèle de planification de séquences de signaux non-verbaux pour l'expression d'attitudes.

Synthèse du chapitre

1. Nous avons introduit un modèle de planification de séquences de signaux non-verbaux pour l'expression d'attitudes. La planification d'une séquence comprend trois étapes : construction de *séquences minimales*, génération de *séquences candidates* et sélection de la séquence finale.
2. Nous avons réalisé une étude en ligne pour vérifier que le modèle est en mesure d'exprimer les attitudes voulues et avec la bonne intensité. Les résultats de l'étude ont montré que les augmentations d'amicalité, baisses d'amicalité et augmentation de dominance étaient reconnues. Cependant, la baisse de dominance n'était pas reconnue par les participants (ceci étant peut-être dû au rôle de recruteur virtuel impliquant un pouvoir sur le candidat). De plus, les intensités des attitudes n'ont pas non plus été reconnues.

7 Implémentation et évaluation d'un recruteur virtuel autonome exprimant des attitudes sociales

Lors d'entretiens d'embauche, un candidat peut être confronté à un recruteur exprimant différentes attitudes sociales. Il peut se montrer plus ou moins amical, et plus ou moins dominant. Une des difficultés dans un entretien d'embauche peut résider dans la capacité pour le candidat à faire face à ces différentes attitudes et à s'y adapter. Par conséquent, un Agent Conversationnel Animé développé pour entrainer des personnes à passer des entretiens d'embauches doit pouvoir exprimer de telles attitudes afin d'y confronter les utilisateurs. Cependant, les recruteurs virtuels utilisés dans les systèmes d'entrainement aux entretiens d'embauche existants n'ont jusqu'ici pas considéré l'expression d'attitudes sociales par le recruteur (voir Section 3.1.2). Le modèle de planification de séquences de signaux non-verbaux que nous avons présenté dans le chapitre 6 permet de doter un Agent Conversationnel Animé de la capacité d'exprimer différentes attitudes. A partir de ce modèle, nous avons implémenté un agent prenant le rôle d'un recruteur virtuel. Ce recruteur a ensuite été évalué par le biais d'une étude où des participants prenaient part à un entretien d'embauche virtuel.

Ce chapitre est organisé en trois sections. Dans une première section, l'implémentation du recruteur virtuel est présentée. Ce dernier est intégré à la plate-forme de simulation d'entretiens d'embauche élaborée dans le projet Tardis. La deuxième section présente l'étude que nous avons réalisée afin d'évaluer si les attitudes exprimées par le recruteur virtuel sont bien identifiées. Nous comparons la perception des attitudes exprimées entre un groupe de participants interagissant directement avec le personnage virtuel et un groupe regardant des vidéos de celui-ci. Enfin, dans la troisième section, nous analysons les résultats de notre étude.

7.1 Implémentation d'un recruteur virtuel pour la simulation d'entretiens d'embauche

Nous avons participé à la conception de l'architecture globale de la plate-forme Tardis et à son implémentation. Le but de cette plate-forme est de permettre à un utilisateur de s'entraîner à passer des entretiens d'embauches. Le modèle de planification de séquences de signaux non-verbaux pour l'expression d'attitudes sociales que nous avons présenté dans le chapitre 6 a été intégré dans cette plateforme. La figure 7.1 est une représentation simplifiée de la plate-forme Tardis montrant ses principaux composants, que nous présentons dans la sous-section suivante.

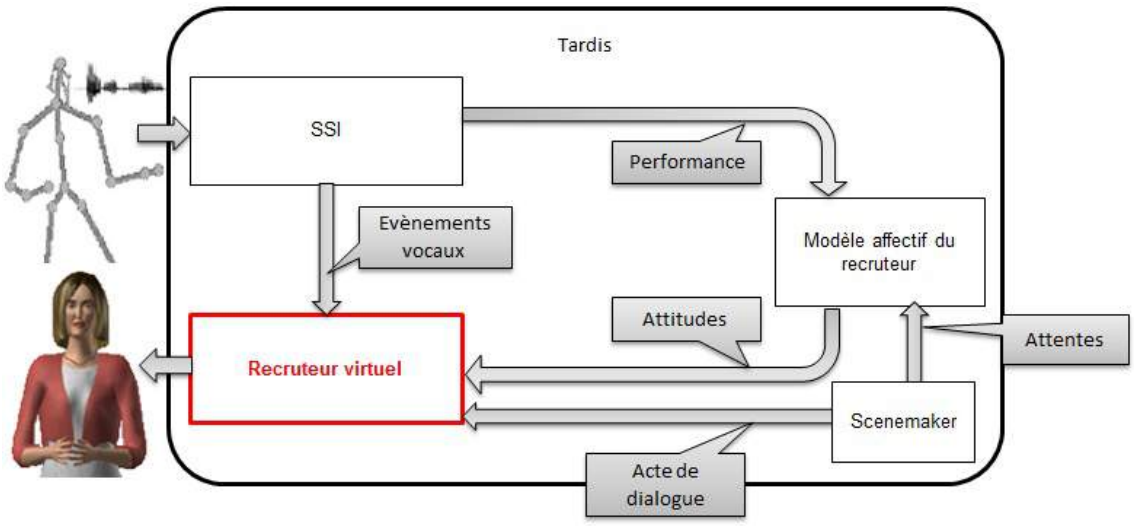


FIGURE 7.1: Architecture de la plate-forme Tardis. En rouge, le recruteur virtuel que nous présentons dans ce chapitre.

7.1.1 Composants du système Tardis

Le plate-forme de simulation d'entretiens d'embauche Tardis est composée de quatre composants principaux :

- Scenemaker [Gebhard et al., 2012] est un logiciel permettant de créer et exécuter des modèles de scénario. La structure d'un entretien d'embauche y est représentée. Celle-ci est constituée de plusieurs phases (*ex.* salutations, introduction, discussion sur les compétences, *etc.*). Chaque phase est constituée

d'un ensemble de types de phrases et de questions que le recruteur peut choisir. Chacun de ces types de phrases est représenté par un *acte de dialogue*, représentant l'objectif de la phrase indépendamment du choix précis des mots qui la réalisent (*ex. ASK_SKILL_TEAMWORK* représente l'action de demander au participant s'il peut travailler en équipe).

- Le logiciel SSI (*Social Signal Interpretation*) [Wagner et al., 2013] est un système de reconnaissance et d'interprétation de signaux sociaux. Il permet la reconnaissance de gestes, postures, expressions faciales (par le biais d'une webcam et d'une Microsoft Kinect) et évènements vocaux (*ex.* baisse d'énergie, augmentation de la hauteur de la voix, présence ou absence de voix). Les évènements vocaux détectés par SSI sont directement utilisés par le recruteur virtuel afin d'assurer le tour de parole et le comportement d'écoute (voir Section 7.1.2.1). SSI calcule de plus une valeur de *performance* de l'utilisateur dans sa réponse à une question, à partir d'éléments vocaux comme la durée de la réponse (une réponse plus longue indique une bonne performance) ou la variation de la prosodie (une voix monotone est moins bien notée), et d'éléments non-verbaux, comme la quantité de gestes ou certaines postures. Cette *performance* est envoyée au modèle affectif du recruteur.

- Le modèle affectif du recruteur a pour rôle de calculer les attitudes que ce dernier doit exprimer, en fonction de la performance de l'utilisateur par rapport aux questions qui lui sont posées [Jones et al., 2014]. Pour chaque question du scénario est définie une valeur de performance *attendue* : par exemple, une bonne performance est attendue pour une question facile. Inversement, on attend des réponses de moins bonne qualité lorsque la question du recruteur est difficile. Le modèle compare les performances observées de l'utilisateur par rapport aux performances attendues, et en détermine des émotions ressenties par le recruteur. Par exemple, une bonne performance à une question difficile fait ressentir de la joie au recruteur. Ces émotions influencent ensuite sur l'humeur du recruteur (des émotions positives améliorent son humeur), qui influent ensuite sur son attitude (une bonne humeur améliore l'attitude du recruteur envers l'utilisateur). A chaque nouvelle question, une nouvelle valeur d'attitude est ainsi calculée. La variation d'attitude par rapport à l'attitude précédente est envoyée au recruteur virtuel. Cependant, le modèle affectif n'envoie pas d'attitudes de soumission, car l'hypothèse est faite que le recruteur

est forcément dominant.

Il peut être noté que la plate-forme Tardis ne contient pas de logiciel de reconnaissance de parole. Pourtant, le comportement verbal d'un candidat (*i.e.* choix des mots) et ses réponses aux questions du recruteur sont certainement très importantes lors d'un entretien d'embauche. Cependant, évaluer la qualité d'une réponse (ouverte) d'un utilisateur à une question du recruteur virtuel serait une tâche difficile. Dans un premier temps, le choix a donc été fait de ne considérer que le comportement non-verbal et vocal des utilisateurs. Le dernier module est le recruteur virtuel que nous avons développé et intégré dans cette plate-forme. Nous présentons son implémentation dans la section suivante.

7.1.2 Implémentation d'un recruteur virtuel

L'architecture du recruteur virtuel que nous intégrons dans la plate-forme Tardis est représentée dans la figure 7.2. Cette architecture est centrée autour du modèle de planification de séquences que nous avons présenté au chapitre 6. Ce modèle va permettre de doter le recruteur virtuel de la capacité d'exprimer des attitudes sociales à travers son comportement non-verbal. Cependant, notre modèle assure uniquement la planification du comportement non-verbal lors des prises de parole du recruteur. Nous l'avons donc complété par d'autres composants afin d'assurer les autres fonctions de la communication multimodale.

Voici une vue d'ensemble du déroulement d'une simulation d'entretien d'embauche avec le recruteur virtuel. Au début d'un entretien, c'est le recruteur virtuel qui démarre l'interaction en posant une première question à l'utilisateur. A partir de là, l'entretien consiste en une série de boucles d'interaction. Une boucle d'interaction commence au moment où le recruteur a fini une phrase et donne la parole à l'utilisateur. Pendant que l'utilisateur répond, le *planificateur de comportement d'écoute* assure la production de *backchannels* et détecte l'instant où le recruteur peut reprendre la parole. Lorsque c'est le cas, celui-ci envoie un message au *moteur de dialogue* qui calcule alors la prochaine phrase que le recruteur virtuel va prononcer. Le *planificateur de séquences* et le *modulateur d'expressivité* calculent alors une séquence de signaux non-verbaux qui va accompagner la phrase. Enfin, cette séquence de signaux non-verbaux est envoyée au *réalisateur de comportements* qui va calculer

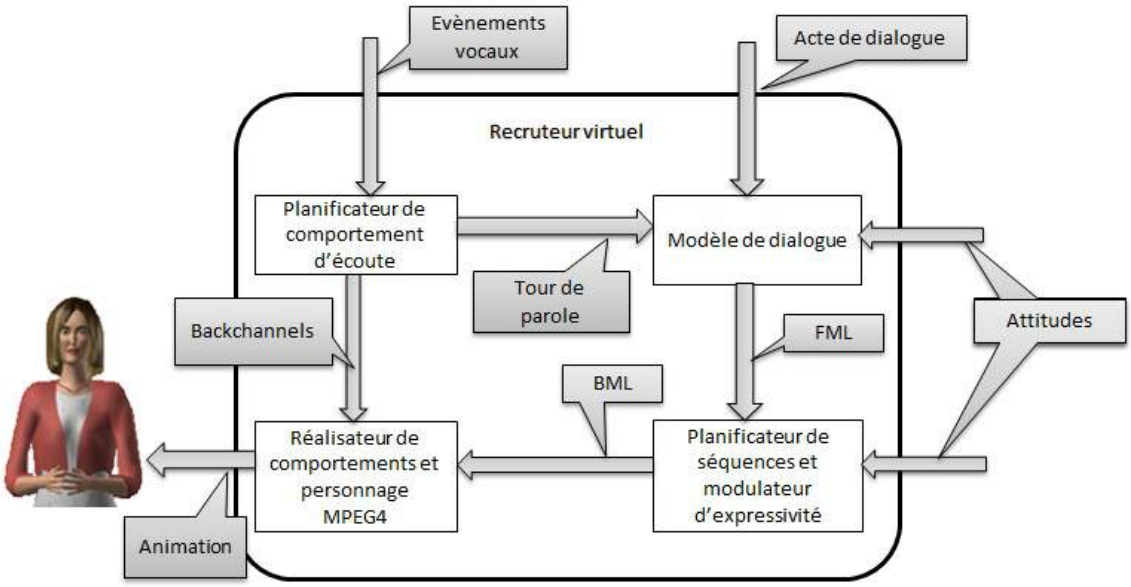


FIGURE 7.2: Représentation de l'architecture du recruteur virtuel.

une animation pour le recruteur virtuel, synchronisée avec le contenu de la nouvelle question. Une fois cette question posée, le cycle recommence avec une nouvelle boucle d'interaction.

Dans les sous-sections suivantes, nous présentons les différents composants du recruteur virtuel, en suivant l'ordre des traitements réalisés dans une boucle d'interaction.

7.1.2.1 Planificateur de comportement d'écoute

Le *Planificateur de comportement d'écoute* remplit deux fonctions principales. Sa première fonction est de générer des signaux non-verbaux de retours lorsque l'utilisateur parle (*backchannels*). Dans le cadre d'interactions avec des ACAs, ceci permet en effet d'améliorer l'engagement des utilisateurs et d'obtenir des phrases plus longues de leur part [Gratch et al., 2006]. Sa deuxième fonction est d'assurer un tour de parole approprié en détectant quand l'utilisateur a fini de parler, le recruteur virtuel pouvant alors poser sa prochaine question.

Ce composant reçoit en entrée des messages décrivant des évènements vocaux détectés dans le signal audio de la voix de l'utilisateur. Ces évènements sont détectés par le logiciel SSI [Wagner et al., 2013]. Par exemple, un message va être envoyé lorsque

l'intensité de la voix de l'utilisateur augmente, lorsqu'une intonation montante ou descendante est détectée, ou encore lorsque l'utilisateur ne parle plus. Ces évènements vocaux sont utilisés pour guider le comportement d'écoute (*i.e.* la production de *backchannels*) et de prise du tour de parole du recruteur virtuel.

Pour la production de *backchannels*, nous avons intégré le *Listener Intent Planner* proposé par Bevacqua *et al.* dans le cadre du projet Semaine [Bevacqua et al., 2012]. A partir d'évènements audio, ce modèle déclenche des *backchannels* et imitations d'expressions faciales (*mimicry*) dont la fréquence et le type sont adaptés en fonction d'une personnalité choisie. Dans notre travail, nous n'avons pas considéré l'influence de l'attitude que doit exprimer le recruteur virtuel sur son comportement d'écoute. Nous développons nos perspectives de travaux futurs sur cette limite de notre modèle de recruteur virtuel dans la section 9.3. Nous avons utilisé les résultats obtenus précédemment par Bevacqua *et al.* sur la perception de *backchannels* [Bevacqua et al., 2010] afin que le comportement d'écoute du recruteur virtuel soit le plus neutre possible. Nous choisissons de sélectionner uniquement des *backchannels* consistant en de légers hochements de tête ou mouvements de sourcils, qui sont considérés comme les plus neutres.

Pour le tour de parole, nous avons utilisé un modèle développé par Gebhard *et al.* pour la plate-forme Tardis [Gebhard et al., 2014]. Celui-ci consiste en un automate à états finis représenté sur la figure 7.3. Son fonctionnement est le suivant : lorsque le recruteur a terminé sa phrase (R_Give_Turn), il y a deux possibilités. Soit l'utilisateur reste silencieux pendant un long moment, et l'agent va proposer une nouvelle question à l'utilisateur ($U\_Timeout$), soit on va détecter une prise de parole de l'utilisateur (U_VA pour $User\ Voice\ Activation$). Une fois que l'utilisateur arrête de parler ($U\_VA = False$), on vérifie que celui-ci ne veuille pas reprendre le tour de parole (U_Check) rapidement dans les secondes à venir (*ex.* si l'utilisateur fait simplement une pause dans sa phrase, alors l'interrompre pourrait le perturber). S'il reprend la parole, alors on revient dans cet état une fois qu'il a terminé ce nouveau tour de parole (U_Again, puis U_Check à nouveau). Si ce n'est pas le cas, alors l'agent reprend la parole (R_Take_Turn), et envoie un message au composant du *Modèle de dialogue* indiquant que la prochaine question peut être posée.

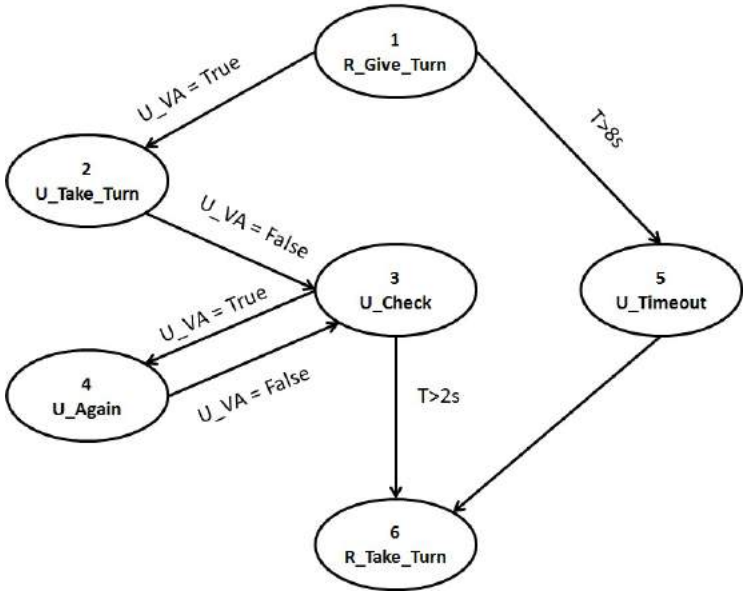


FIGURE 7.3: Automate à états finis du modèle de tour de parole de [Gebhard et al., 2014].

7.1.2.2 Modèle de dialogue

Lorsque le *Modèle de dialogue* reçoit un message indiquant que le recruteur peut prendre la parole, celui-ci interroge le modèle de scénario (Scenemaker, présenté dans la section 7.1.1) qui fournit le prochain acte de dialogue que le recruteur doit effectuer. Le rôle du *Modèle de dialogue* est de transformer un acte de dialogue en un fichier FML contenant une phrase enrichie d'intentions communicatives (voir Section 6.1).

Le choix des mots pour réaliser un acte de dialogue donné peut influencer l'attitude exprimée [Bee et al., 2010]. Nous avons donc choisi d'utiliser un modèle de dialogue considérant les attitudes sociales. Pour cela, nous avons intégré le modèle de Callejas *et al.* [Callejas et al., 2014]. Les auteurs ont défini une collection d'actes de dialogue utilisés dans des entretiens d'embauche. Chaque acte de dialogue est associé à plusieurs fichiers FML, un fichier contenant une phrase exprimant une attitude neutre, un exprimant une attitude amicale, et un dernier exprimant une attitude hostile. Les phrases de ces fichiers FML ont été rédigées à partir des règles issues du travail de Mairesse *et al.* dans le projet PERSONAGE [Mairesse & Walker, 2011]. Des aspects comme la longueur des phrases, la variété du vocabulaire ou la quantité de pronoms

utilisés sont pris en compte afin d'adapter l'attitude d'une phrase pour le même acte de dialogue. Les auteurs ont montré que leur modèle parvenait à exprimer des attitudes lorsqu'il était joint avec un modèle de comportement non-verbal [Callejas et al., 2014].

Nous avons étendu cet ensemble de fichiers FML pour inclure des phrases exprimant des attitudes dominantes. Pour cela nous avons utilisé le modèle de Linssen *et al.* afin d'écrire de nouvelles variations de phrases pour la dominance [Linssen et al., 2014]. Linssen *et al.* lient différentes zones du circomplexe interpersonnel (*i.e.* différentes combinaisons d'amicalité et de dominance) à des stratégies de politesse issues des travaux de Brown et Levinson et de Culpeper *et al.* [Brown, 1987, Culpeper et al., 2003]. Par exemple, pour exprimer une attitude sur l'axe d'amicalité et de soumission, Linssen *et al.* proposent d'utiliser des stratégies de « politesse positive » (Prendre la volonté de son interlocuteur en considération : *ex.* « Voulez-vous bien me passer ce livre ? ») et de « politesse négative » (Ne pas empiéter sur l'autonomie de son interlocuteur : *ex.* « Si cela ne vous dérange pas, pouvez-vous me passer ce livre ? »).

Nous n'avons pas créé de versions des actes de dialogue pour l'attitude de soumission, car le module affectif envoyant les attitudes à exprimer au planificateur de séquences dans le projet Tardis ne considère pas cette attitude. Voici un exemple des phrases utilisées pour exprimer une attitude lors de l'expression de l'acte de dialogue *Ask_Academic_Interest*, dont le but est de demander à l'utilisateur pourquoi il a fait des études dans sa spécialité.

Neutre - Pourquoi avez-vous fait des études dans votre spécialité ?
Dominante - Maintenant, vous allez me dire pourquoi vous avez fait des études dans votre spécialité.
Amicalité - Qu'est-ce qui vous a plu le plus dans votre spécialité à l'université ?
Inamicalité - Pour quelle raisons avez vous choisi votre spécialité à l'université ?

7.1.2.3 Planificateur de séquences

Le *Planificateur de séquences* est le modèle que nous avons présenté dans le chapitre 6. Son rôle est de transformer, en fonction d'une attitude donnée en entrée (*ex.* augmentation de dominance, baisse d'amicalité), une phrase annotée avec des inten-

tions communicatives exprimées dans le format FML en une séquence de signaux non-verbaux exprimée dans le format BML.

Après l'étude préliminaire de notre modèle consistant à évaluer si les types et intensités des attitudes exprimées étaient bien reconnues (Section 6.3), nous l'avons modifié afin de ne plus considérer les différentes intensités de variations d'attitude, celles-ci n'étant pas du tout reconnues. Ainsi, lors de la génération de séquences pour une variation d'attitude donnée, par exemple une augmentation de dominance, l'algorithme sélectionne la meilleure séquence en utilisant toutes les séquences annotées d'augmentation de dominance, qu'elles soient fortes ou faibles.

Comme nous l'avons présenté en section 2.2.3, l'amplitude des gestes et leur intensité peuvent affecter la perception de l'attitude exprimée. Lorsqu'une séquence de signaux non-verbaux est calculée par notre modèle de planification de séquences, le message BML qui l'exprime est alors envoyé au prochain composant, le *Modulateur d'expressivité*, qui va ajuster l'amplitude et l'intensité des gestes en fonction de l'attitude à exprimer.

7.1.2.4 Modulateur d'expressivité

Le rôle du *Modulateur d'expressivité* est d'ajuster certains paramètres de réalisation de gestes. Nous transformons les gestes pour les rendre plus ou moins amples et plus ou moins intenses en fonction de l'attitude à exprimer. Pour cela, nous utilisons la représentation des paramètres d'expressivité utilisée dans la plate-forme d'Agent Conversationnel Greta [Hartmann et al., 2006]. Les définitions de gestes, de mouvements de tête et de torse du format BML y sont étendues par des paramètres d'expressivité. Par exemple, les amplitudes de gestes, de hochements de tête ou de changements de posture y sont leurs intensités sont représentées par le paramètre PWR (pour *Power*, puissance). Par exemple, un geste défini avec un paramètre $SPC = 0.7$ sera plutôt ample, tandis qu'un geste avec un paramètre $PWR = 0.1$ sera très peu intense. La figure 7.4 montre une définition de geste dans le format BML étendu pour représenter des paramètres d'expressivité.

Pour créer le modulateur d'expressivité, nous avons utilisé le modèle de Ravenet *et al.* [Ravenet et al., 2013], que nous avons présenté en section 3.2. Pour différentes intentions communicatives, une base de données d'expressions multimodales

créées par des utilisateurs a été récoltée par une méthode de *crowdsourcing*. Lors de la création de ces expressions, les utilisateurs pouvaient sélectionner l'amplitude et l'intensité des gestes. Ces données ont été utilisées pour créer un réseau Bayésien modélisant l'influence d'une attitude et d'une intention sur le choix des signaux non-verbaux d'un Agent Conversationnel Animé et sur les paramètres d'expressivité de ces signaux. Ce modèle a été évalué dans le cadre de générations d'expressions multimodales conjointement à un modèle de dialogue pour l'expression de l'amicalité et de l'inamicalité [Callejas et al., 2014]. Dans notre cas, lorsque le modèle de planification de séquences sélectionne un geste, le modulateur d'expressivité interroge le réseau Bayésien et obtient ainsi des paramètres d'expressivité SPC et PWR appropriés.


```
<?xml version="1.0"?>
<bml xmlns="http://www.mindmakers.org/projects/BML" character="Greta"
    id="bml2">
    <gesture ready="0.0" relax="2.260" id="g1" lexeme="ask_Ges_B">
        <description type="gretabml">
            <reference>performative=ask_Ges_B</reference>
            <SPC>0.396</SPC>
            <TMP>0.131</TMP>
            <FLD>0.500</FLD>
            <PWR>0.520</PWR>
            <REP>0.000</REP>
            <OPN>0.610</OPN>
            <TEN>0.000</TEN>
        </description>
    </gesture>
</bml>
```

FIGURE 7.4: Définition de geste dans le format BML étendu pour la représentation de paramètres d'expressivité. Les paramètres d'expressivité de signaux non-verbaux sont définis dans les balises enfants de la balise `<gesture>`. SPC : amplitude spatiale. TMP : rapidité du geste. FLD : fluidité. PWR : intensité. REP : nombre de répétitions. OPN : ouverture. TEN : tension.

7.1.2.5 Réalisateur de comportements et personnage MPEG-4

Le *Réalisateur de comportements* constitue le bout de la chaîne de notre modèle de recruteur virtuel. Son rôle est de synthétiser l'animation du personnage virtuel.

Nous avons utilisé la plate-forme modulaire de construction d'Agent Conversationnel Animé VIB [Pez et al., 2014] (extension de Greta : une capture d'écran de VIB est présentée en figure 7.5) qui contient un réalisateur de comportement compatible avec la norme BML et avec les paramètres d'expressivité de [Hartmann et al., 2006]. A ce réalisateur de comportement sont associés plusieurs modèles de synthèse d'animation pour les différentes modalités de l'ACA, qui permettent de transformer les différentes parties d'un message BML en des séquences de paramètres d'animation, suivant la norme MPEG-4. Ces animations sont enfin utilisées pour animer un personnage virtuel, compatible avec la norme MPEG-4, intégré dans un environnement 3D affiché par le moteur Ogre.

L'architecture et son implémentation que nous avons présentées ont permis de développer un recruteur virtuel pouvant exprimer différentes attitudes lorsqu'il prend la parole, à la la fois par son comportement verbal et son comportement non-verbal. La prochaine étape consiste à évaluer si les attitudes exprimées par un tel recruteur sont correctement identifiées par un utilisateur en interaction avec celui-ci. Cette évaluation est décrite dans la section suivante.

7.2 ÉVALUATION

L'étude que nous avons présentée dans le chapitre 6 avait pour but d'évaluer si notre modèle de planification de séquences permettait d'exprimer des attitudes. Le modèle utilisé dans l'étude ne comportait pas d'adaptation de l'expressivité, et l'influence du discours du recruteur n'était pas prise en compte. Enfin, nous avions évalué la perception du recruteur virtuel par des personnes regardant des vidéos : les participants n'étaient pas impliqués dans l'interaction. Ils n'avaient pas à penser et produire des réponses aux questions du recruteur. Pour pallier ces limites, nous avons proposé une architecture de recruteur virtuel considérant l'influence de l'expressivité des gestes et du discours sur l'expression de l'attitude, et nous avons réalisé une nouvelle étude afin d'évaluer les attitudes perçues lors d'interactions avec le recruteur virtuel.

Cette étude poursuit plusieurs objectifs. Tout d'abord, les modules ajoutés pour rendre le recruteur virtuel complet et autonome participent à l'expression d'une attitude : le modèle de dialogue et le modulateur d'expressivité ont été conçus pour

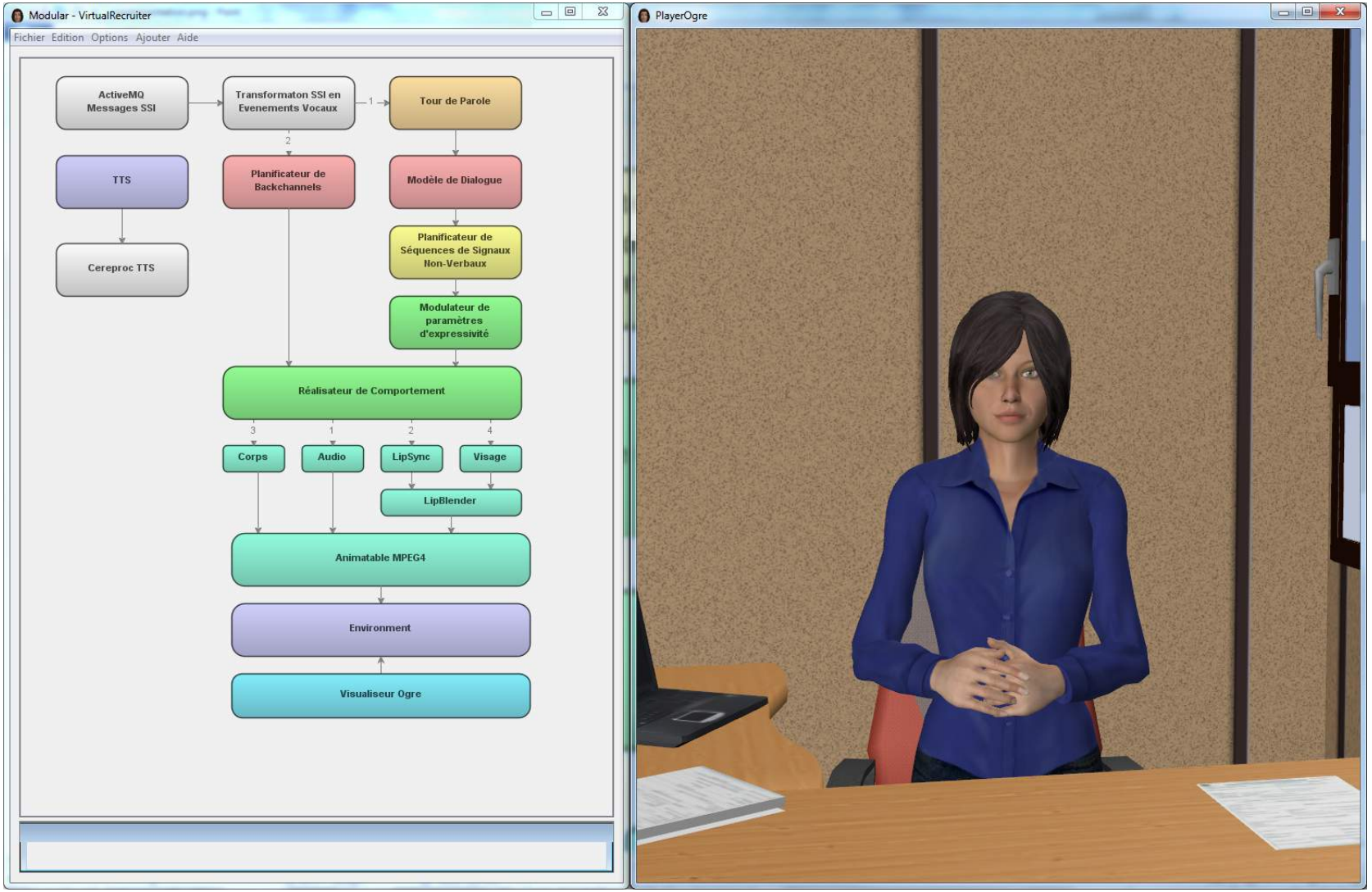


FIGURE 7.5: Implémentation du recruteur virtuel dans le logiciel de construction d'Agent Conversationnel Animé modulaire VIB. A gauche, l'interface de création d'ACA, comprenant la chaîne de modules correspondant à notre implémentation du recruteur virtuel. A droite, le personnage MPEG-4 et l'environnement 3D intégrés intégrés dans le moteur Ogre.

contribuer à l'expression d'une attitude. Nous avons ainsi voulu évaluer la perception des attitudes exprimées par le recruteur pour vérifier que celles-ci sont bien identifiées. De plus, nous avons voulu comparer les contributions respectives du modèle de planification du comportement non-verbal et du modèle de dialogue. Enfin, il s'agit de comparer la perception de l'attitude par le participant selon qu'il interagit avec l'ACA ou bien qu'il regarde des vidéos de ce personnage. Nous nous sommes ainsi posé les trois questions de recherche suivantes lors de la construction de cette étude.

QR1 : Les attitudes exprimées avec le modèle complet (multimodal) sont elles correctement identifiées ?

QR2 : Quelle combinaison de modalités (verbale, non-verbale, multimodale) permet d'atteindre le meilleur taux de reconnaissance des attitudes ?

QR3 : Une différence dans l'identification de l'attitude exprimée existe-t-elle lorsque les expressions de l'agent sont évaluées en interaction par rapport à une évaluation à partir de vidéos uniquement ?

Nous présentons maintenant le protocole expérimental que nous avons adopté pour répondre à ces trois questions.

7.2.1 Protocole expérimental

7.2.1.1 Conditions et variables indépendantes

Nous avons voulu vérifier si les différentes attitudes exprimées par notre modèle étaient bien perçues, lors d'évaluations avec des vidéos ou en interaction, et avec différentes combinaisons de modalités. Nous avons donc défini les variables indépendantes suivantes.

Notre modèle pouvant exprimer plusieurs attitudes, la première variable indépendante VI_1 que nous avons considérée est le type d'attitude exprimé : la dominance, l'amicalité, ou l'inamicalité. Ensuite notre deuxième question de recherche $QR2$ étant dédiée à la comparaison des différentes modalités impliquées dans l'expression de l'attitude, nous avons inclus une variable indépendante VI_2 pour représenter les différentes combinaisons de modalités. L'attitude peut être exprimée par le mo-

Inter \ Intra		Neutre	Dialogue	Non-verbal	Multimodal
Interaction	Dominant	I1	I2	I3	I4
	Amical	I5	I6	I7	I8
	Inamical	I9	I10	I11	I12
Vidéo	Dominant	V1	V2	V3	V4
	Amical	V5	V6	V7	V8
	Inamical	V9	V10	V11	V12

TABLE 7.1: Conditions de l'étude. Chaque participant est réparti sur une ligne. Par exemple, un participant évaluant un agent inamical par des vidéos sera exposé aux conditions V9, V10, V11 et V12.

dèle de dialogue uniquement (*Dialogue*), par le modèle de comportement non-verbal uniquement (*Non-verbal*), par aucun des deux modèles (*Neutre*) ou bien par les deux modèles à la fois (*Multimodal*). Enfin, une troisième variable indépendante VI_3 correspond au mode d'évaluation, qui est étudié dans notre troisième question de recherche $QR3$.

$VI_1 : Attitude \in \{Dominant, Amical, Inamical\}$

$VI_2 : Modalité \in \{Neutre, Dialogue, Nonverbal, Multimodal\}$

$VI_3 : Medium \in \{Interaction, Vidéo\}$

Nous avons choisi d'adopter une répartition des participants avec l'attitude et le mode d'évaluation comme variables indépendantes *inter-participants*, et la modalité comme variable indépendante *intra-participants*. Ainsi, un participant était exposé aux conditions représentées sur une des lignes de la table 7.1. Par exemple, un participant évaluant les attitudes d'amicalité en interaction est exposé aux conditions I5, I6, I7 et I8. L'ordre de présentation de ces attitudes était contrebalancé pour éviter un effet de l'ordre sur les résultats.

7.2.1.2 Déroulement

Afin de comparer la perception des attitudes de notre recruteur virtuel en interaction d'une part et au travers de vidéos d'autre part, nous avons réalisé deux supports pour notre évaluation. Le premier consistait en un système de simulation d'entretiens d'embauche avec lequel des participants pouvaient interagir avec le recruteur, et le

deuxième consistait en une plate-forme web.

Évaluation en interaction :

Nous décrivons d'abord le déroulement de l'évaluation en interaction, puis la salle d'étude que nous avons aménagée pour qu'elle ait lieu, et enfin le système que nous avons développé pour la réaliser.

Avant l'interaction - Les participants commencent par lire et remplir un formulaire de consentement. Il leur est indiqué que leur participation est anonyme et que les seules données conservées sont leurs réponses à des questionnaires : aucune donnée audio-visuelle n'est conservée. De plus, il leur est précisé que les personnages de l'expérience ne sont pas dotés de système de compréhension de la parole. Nous avons choisi d'informer les participants de cela afin qu'ils ne s'attendent pas à ce que le recruteur virtuel les comprenne, et ainsi éviter que l'absence de compréhension de sa part influence leur perception des attitudes exprimées. Le formulaire de consentement est reproduit dans l'annexe A. Pendant ce temps, un ensemble de conditions est assigné aléatoirement au participant (voir Table 7.1), et un ordre d'exposition aux modalités est choisi afin de contrebalancer l'effet de cet ordre. Le système est alors configuré en tenant compte de ces paramètres. Par exemple, un participant pouvait être assigné à l'attitude amicale, c'est à dire les conditions I5, I6, I7 et I8, avec un ordre d'exposition aux modalités *Dialogue*, *Multimodal*, *Non-verbal* et *Neutre*, c'est à dire I6, puis I8, I7 et enfin I5. Les participants lisent ensuite des instructions qui leur expliquent le déroulement de l'expérience (voir Annexe A, page 191), remplissent quelques informations démographiques (âge, nationalité, sexe), et peuvent poser des questions sur le déroulement de l'expérience.

Phase d'habituation - Une fois le participant prêt, la prochaine étape de l'expérience consiste en une phase d'habituation. Le responsable de l'expérience quitte alors la pièce et active l'interaction à distance : le reste de la phase est ensuite automatique. L'écran initialement noir fait apparaître l'environnement virtuel et un premier Agent Conversationnel Animé (Figure 7.6, gauche) qui donne à nouveau les instructions au participant, sous forme de dialogue simple. Cela a pour but de donner l'opportunité au participant de s'habituer à la situation. Ici, aucune attitude n'est exprimée par l'ACA. Cette phase permet aussi de vérifier qu'il n'y a pas de problème technique (*ex.* problème de microphone ou de déclenchement des phrases par l'ACA). Les phrases utilisées dans cette phase sont reproduites dans l'annexe A (Section A.3).

FIGURE 7.6: Les deux personnages utilisés pour réaliser l'étude. A gauche, le personnage réalisant la phase d'habituation. A droite, le personnage prenant le rôle du recruteur virtuel.

Une fois le dialogue terminé, l'écran redevient noir, et une nouvelle opportunité est donnée au participant de poser des questions au responsable de l'expérience. Une fois cela terminé, la phase d'évaluation commence.

Phase d'évaluation - La phase d'évaluation consiste en une simulation d'entretien d'embauche, divisée en quatre parties. Afin que les participants ne soient pas influencés par l'impression que leur a laissé le personnage de la phrase d'habituation, nous avons utilisé un deuxième modèle de personnage virtuel pour prendre le rôle du recruteur (voir Figure 7.6, droite). Dans chacune des parties, le recruteur pose des questions au participant sur un thème particulier. Dans la première partie, il s'agit de questions d'ordre général. Dans la deuxième partie, la formation et les études du participant sont évoquées. Ensuite, le recruteur l'interroge sur son expérience professionnelle. Enfin, la dernière partie concerne les compétences interpersonnelles du participant. Nous avons choisi ces thèmes car ceux-ci correspondent à des sujets régulièrement évoqués par les recruteurs de notre corpus d'entretiens d'embauche. Les actes de dialogue et phrases correspondantes du modèle de dialogue sont détaillées dans la section A.3 de l'annexe A. Pour chaque partie, le recruteur virtuel commence par indiquer son thème, puis pose une série de trois questions auxquelles

FIGURE 7.7: La salle d'étude de l'évaluation du recruteur en interaction.

le participant répond. Ensuite, il remercie le participant et lui demande de remplir le questionnaire sur l'ordinateur placé sur le bureau. L'écran redevient alors noir pour indiquer que l'interaction est mise en pause. Ce questionnaire est détaillé dans la section 7.2.1.3, et la page web correspondante est reproduite dans l'annexe A. Une fois le questionnaire rempli par le participant, une requête est envoyée du navigateur qui va déclencher la prochaine phase. Ainsi, de la même manière que pour la phase d'habituation, c'est par une commande à distance que la phase d'évaluation est lancée, mais l'interaction est ensuite totalement automatisée. A la fin de ces quatre phases, les participants sont remerciés.

Salle d'étude - Une salle d'étude a été aménagée pour la réalisation de l'évaluation en interaction. Pour améliorer l'immersion des participants, nous avons utilisé un grand écran permettant d'afficher un agent à taille réelle (une photographie de la salle d'étude est présentée en figure 7.7). La scène 3D présente le personnage virtuel assis derrière un bureau. Un bureau était placé devant l'écran, au niveau du bureau de l'environnement virtuel afin de donner une impression de continuité. Sur ce bureau étaient disposés un microphone, utilisé par les participants pendant l'interaction, et un ordinateur, pour récolter les impressions des participants au fur et à mesure de l'interaction (voir Section 7.2.1.3).

Plate-forme utilisée pour l'évaluation - Nous avons utilisé une version simplifiée de la plate-forme de simulation d'entretiens d'embauche Tardis pour notre évaluation (voir Figure 7.2). Le logiciel SSI [Wagner et al., 2013] est utilisé afin de récupérer des évènements audio (*ex.* activation de la voix, changements dans la prosodie) nécessaires au module de planification du comportement d'écoute et du tour de parole. Le composant d'exécution de scénario est remplacé. Le scénario étant fixé à l'avance, suivant les phases et parties que avons décrites dans les paragraphes précédents, nous avons réalisé un module *ad-hoc* pour envoyer les actes de dialogue correspondant aux trois questions de la partie en cours dans l'ordre voulu. Le modèle affectif n'est pas utilisé, car les attitudes exprimées dépendent de la condition assignée au participant. Enfin, un second module *ad-hoc* se charge de contrôler la combinaison des modalités utilisée pour chaque partie, selon la condition dans laquelle se trouve le participant.

Évaluation par vidéos :

Pour collecter les données des participants évaluant le recruteur virtuel au moyen de vidéos, nous avons créé une plate-forme web. Le participant se connectant à la plate-forme était d'abord accueilli par une première page présentant des instructions (reproduites en annexe A, page 192). Lors de l'arrivée sur cette page, un script se chargeait d'assigner aléatoirement une attitude et un ordre d'affichage des modalités au participant, de la même manière que pour l'évaluation en interaction (lignes de la catégorie « Vidéo » de la table 7.1). Dans l'évaluation par vidéos, les participants n'interagissant pas avec le recruteur virtuel, il n'y avait pas lieu d'inclure une phase d'habituation. Ainsi, une fois les instructions lues et quelques informations démographiques (âge, nationalité, sexe) fournies, les participants passaient directement à la phase d'évaluation.

L'étude par vidéos était divisée en quatre parties abordant les même thèmes que dans l'étude en interaction : questions générales, formation, expérience professionnelle, compétences interpersonnelles. Pour chacune de ces parties, les participants était dirigés vers une page web où ils pouvaient voir trois vidéos correspondant aux trois questions posées par le recruteur virtuel dans cette partie. Pour une question, une attitude et une combinaison de modalités données, ces vidéos correspondaient aux mêmes animations qu'aurait produit le recruteur virtuel lors de l'évaluation en interaction. Le questionnaire était aussi identique à celui utilisé dans l'évaluation en

interaction (voir Section 7.2.1.3).

7.2.1.3 Mesures

Afin d'évaluer l'attitude perçue par les participants lors de nos deux évaluations parallèles, nous avons choisi d'utiliser des mesures de trois types : des échelles avec lesquelles les participants indiquaient directement les degrés de dominance et d'amicalité qu'ils avaient perçus, des échelles construites à partir d'adjectifs subjectifs (inspirées des échelles d'adjectifs utilisées dans [Fukayama et al., 2002]), et différentes échelles évaluant les impressions des participants de manière indirectes. Une échelle de Likert à sept points était utilisée pour chacune de ces questions, reproduites ci-dessous.

Le recruteur virtuel se comporte d'une manière :

Q1 -	Pas du tout dominante	1	2	3	4	5	6	7	Tout à fait dominante
Q2 -	Pas du tout amicale	1	2	3	4	5	6	7	Tout à fait amicale

Le recruteur vous a semblé :

Q3 -	Fermé	1	2	3	4	5	6	7	Ouvert
Q4 -	Froid	1	2	3	4	5	6	7	Chaleureux
Q5 -	Timide	1	2	3	4	5	6	7	Confiant
Q6 -	Effacé	1	2	3	4	5	6	7	Affirmé

Ce recruteur semble favorable au candidat :

Q7 -	Pas du tout	1	2	3	4	5	6	7	Tout à fait

Ce recruteur semble vouloir faire échouer le candidat : :

Q8 -	Pas du tout	1	2	3	4	5	6	7	Tout à fait

Ce recruteur veut mettre à l'aise le candidat :

Q9 -	Pas du tout	1	2	3	4	5	6	7	Tout à fait

Ce recruteur veut déstabiliser le candidat :

Q10 -	Pas du tout	1	2	3	4	5	6	7	Tout à fait

7.2.2 Résultats

Nous avons réuni 24 participants pour participer à l'étude en interaction, et collecté les réponses de 24 participants pour l'étude en ligne. Parmi les 48 participants, 28 étaient des femmes et 20 des hommes. Les participants étaient majoritairement de nationalité française (91, 7%), et l'âge moyen des participants était de 34, 4 ans ($\sigma = 13, 3$). Des tables rassemblant les moyennes des réponses aux dix questions présentées plus haut sont rapportées en annexe A, pages 199-202. Le premier tableau regroupe les réponses des participants ayant interagi avec le recruteur virtuel, le deuxième regroupe les participants ayant noté les vidéos, et le troisième regroupe l'ensemble des participants. Nous présentons maintenant les tests que nous avons réalisés afin de répondre à nos questions de recherche.

Pour pouvoir analyser les taux d'identification correcte des attitudes, nous avons commencé par transformer les données. En effet, dans le cas d'une séquence exprimant l'amicalité, l'attitude est identifiée si un score plus haut que quatre est indiqué à la question $Q2$. Pour l'inamicalité, il faut avoir indiqué un score plus bas que quatre à $Q2$, et pour la dominance, un score plus haut que quatre à la question $Q1$. Nous appliquons le même procédé aux échelles d'adjectifs ($Q3 - Q6$). Nous transformons donc les données des réponses des participants de la manière suivante, et obtenons ainsi pour chaque attitude une mesure Q_{Obj} évaluant directement l'identification de l'attitude, et deux mesures évaluant l'identification de l'attitude par des échelles d'adjectifs, Q_{Adj1} et Q_{Adj2} :

$$Q_{Obj} = \begin{cases} Q2 - 4 & \text{si Amicalité} \\ 4 - Q2 & \text{si Inamicalité} \\ Q1 - 4 & \text{si Dominance} \end{cases} \qquad Q_{Adj1} = \begin{cases} Q3 - 4 & \text{si Amicalité} \\ 4 - Q3 & \text{si Inamicalité} \\ Q5 - 4 & \text{si Dominance} \end{cases}$$

$$Q_{Adj2} = \begin{cases} Q4 - 4 & \text{si Amicalité} \\ 4 - Q4 & \text{si Inamicalité} \\ Q6 - 4 & \text{si Dominance} \end{cases}$$

Les valeurs de Q_{Obj}, Q_{Adj1} et Q_{Adj2} sont donc des nombres compris entre -3 et 3, une valeur positive indiquant une identification correcte de l'attitude. Les résultats pour Q_{Obj} sont présentés en figure 7.8.

QR1 : Reconnaissance des attitudes avec le modèle complet :

Tout d'abord, nous avons voulu vérifier que les attitudes exprimées avec le modèle

complet, combinant une expression de l'attitude par le modèle de dialogue et le comportement non-verbal du recruteur, étaient bien identifiées. Pour tester cette hypothèse, nous réalisons des tests de Student uni-latéraux entre les scores des attitudes reconnues dans la condition multimodale (MM) et une moyenne $\mu_0 = 0$. Ces tests nous permettent de vérifier que les moyennes des scores de reconnaissance des attitudes sont significativement différentes de la moyenne de l'échelle de mesure.

En considérant toutes les attitudes simultanément, les tests de Student sont significatifs pour Q_{Obj} ($\mu = 0.68, \sigma = 1.18, t(47) = 4.22, p = 0.000$), Q_{Adj1} ($\mu = 0.82, \sigma = 1.43, t(47) = 4.2, p = 0.000$) et Q_{Adj2} ($\mu = 0.95, \sigma = 1.28, t(47) = 5.38, p = 0.000$). Ceci indique qu'en général, les attitudes sont reconnues.

Dans le cas de la dominance seule, les trois mesures sont elles aussi significatives : Q_{Obj} ($\mu = 1.02, \sigma = 0.94, t(15) = 4.70, p = 0.000$), Q_{Adj1} ($\mu = 1.40, \sigma = 1.17, t(15) = 5.29, p = 0.000$) et Q_{Adj2} ($\mu = 1.48, \sigma = 1.47, t(15) = 4.37, p = 0.000$). La dominance semble donc être reconnue.

Pour l'inamicalité considérée seule, Q_{Obj} ($\mu = 0.16, \sigma = 1.20, t(15) = 1.09, p = 0.29$) et Q_{Adj1} ($\mu = 0.10, \sigma = 1.42, t(15) = 0.89, p = 0.39$) ne sont pas significatifs, mais Q_{Adj2} l'est ($\mu = 0.625, \sigma = 1.13, t(15) = 2.57, p = 0.02$). On ne peut donc pas conclure que l'inamicalité soit directement reconnue, cependant les participants jugent le recruteur exprimant de l'inamicalité comme plus froid.

Enfin, pour l'amicalité, les trois mesures sont significatives : Q_{Obj} ($\mu = 0.85, \sigma = 1.25, t(15) = 3.08, p = 0.008$), Q_{Adj1} ($\mu = 0.92, \sigma = 1.42, t(15) = 2.92, p = 0.01$) et Q_{Adj2} ($\mu = 0.77, \sigma = 1.13, t(15) = 3.08, p = 0.008$). L'amicalité semble donc être reconnue.

Pour les questions $Q7 - Q10$, nous comparons les perceptions des différentes attitudes entre elles par le biais d'ANOVAs. Une différence significative est observée pour $Q10$ (« *Ce recruteur veut déstabiliser le candidat* » $F(2, 42) = 2.42, p = 0.02$), le recruteur exprimant de la dominance étant vu comme voulant plus déstabiliser le candidat ($\mu = 3.67$) que lorsqu'il exprime de l'amicalité ($\mu = 2.27$) ou de l'inamicalité ($\mu = 2.48$).

QR2 : Combinaison de modalités :

Notre deuxième question de recherche visait à vérifier quelle combinaison de mo-

dalités permettait une meilleure reconnaissance de l'attitude exprimée. Nous avons réalisé une ANOVA entre les quatre différentes combinaisons de modalités (*Aucune, Dialogue, Non-verbal, Multimodal*) sur les scores combinés de toutes les attitudes possibles (*Amicalité, Inamicalité, Dominance*) et pour les deux modes d'évaluations (*Interaction, Vidéos*).

Pour Q_{Adj1} ($F(3, 188) = 2.28, p = 0.08$) et Q_{Adj2} ($F(3, 188) = 2.36, p = 0.07$), les résultats approchent mais n'atteignent pas le seuil de signifiance $p < 0.05$. Nous remarquons toutefois une tendance à de meilleurs scores dans le cas multimodal (Q_{Adj1} plus fort de 0.56 et Q_{Adj2} plus fort de 0.46 en moyenne dans le cas multimodal). Les attitudes semblent donc être mieux reconnues dans le cas multimodal même si cela n'est pas significatif.

Nous observons un effet significatif du type de combinaison des modalités pour Q_{Obj} ($F(3, 188) = 4.69, p = 0.003$), où le cas multimodal obtient les meilleurs scores (Q_{Obj} plus élevé de 0.71 en moyenne dans le cas multimodal). Les attitudes d'amicalité ($\mu = 0.875$) et de dominance ($\mu = 1$) sont les mieux reconnues. L'inamicalité est moins bien reconnue ($\mu = 0.1875$), mais la différence avec les autres conditions est nette (voir Figure 7.8, bas).

QR3 : Mode d'évaluation :

Notre dernière question de recherche consistait à comparer deux modes d'évaluation du recruteur virtuel : d'une part, les impressions des participants ayant interagi directement avec lui, d'autre part, les participants ayant seulement regardé des vidéos. Nous avons regroupé les mesures indépendamment de l'attitude et de la combinaison de modalités (hormis la condition *Neutre*, que nous avons retirée) et réalisé une ANOVA pour chacune des questions. Une différence significative est effectivement observée pour Q_{Obj} ($F(1, 142) = 4.30, p = 0.039$). Le score de reconnaissance est plus élevé dans le cas des vidéos (Q_{Obj} plus élevé de 0.42 en moyenne). Parmi les questions Q_{Adj1}, Q_{Adj} et de $Q7$ à $Q10$, seule la question $Q10$ (« *Ce recruteur veut déstabiliser le candidat* ») révèle une différence significative dans le cas de l'expression de l'amicalité avec le modèle complet (multimodal), les participants évaluant des vidéos attribuant une plus forte volonté du recruteur à déstabiliser le candidat ($\mu = 2.80$) que dans le cas interactif ($\mu = 1.75$; $F(1, 14) = 5.98, p = 0.03$).

Nous avons voulu comparer cet effet selon la combinaison de modalités employées. A cet effet, nous avons réalisé quatre nouvelles ANOVA sur l'ensemble des réponses,

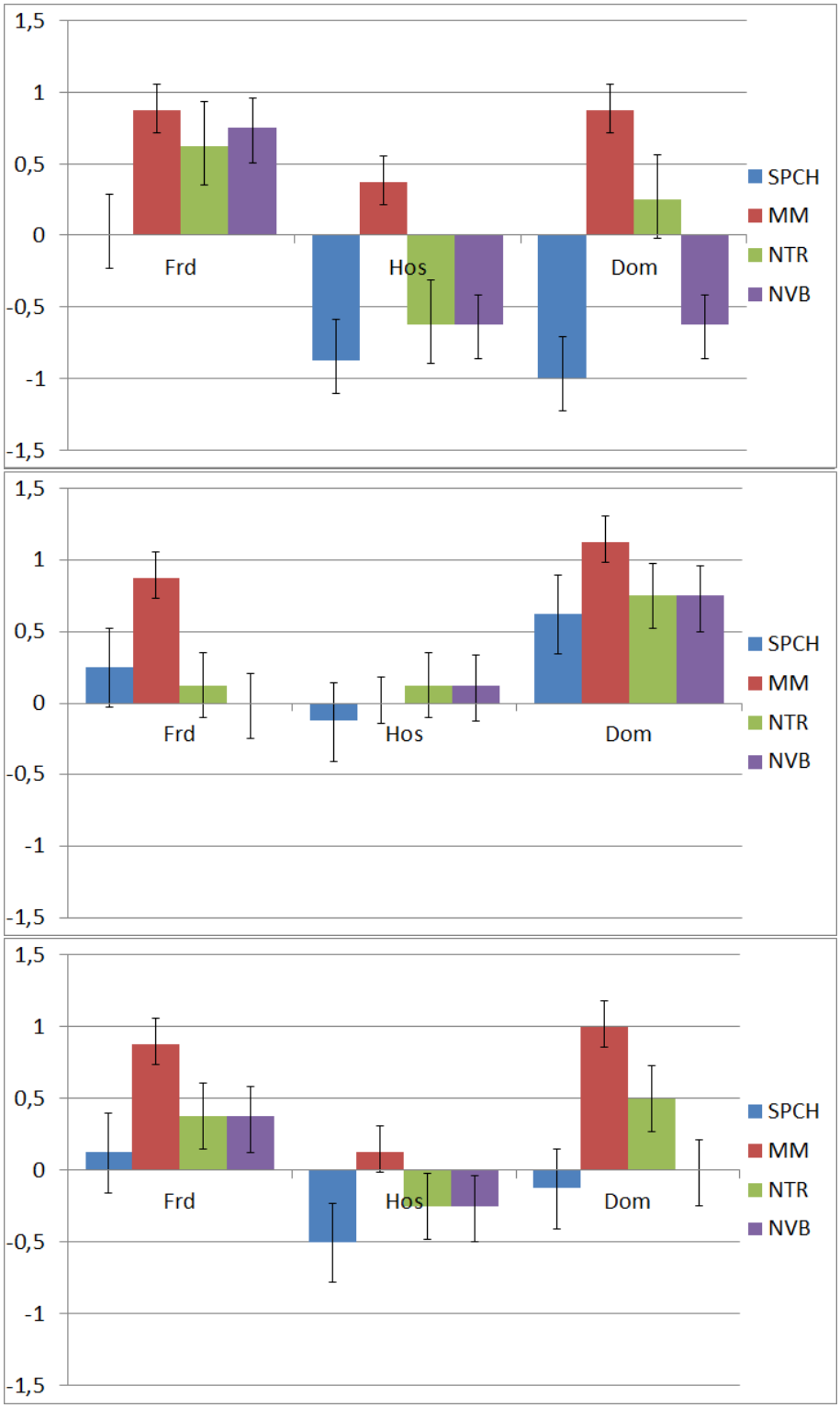


FIGURE 7.8: Score Q_{Obj} par condition. En haut, scores pour les participants en interaction. Au milieu, scores pour les participants par vidéos. En bas, scores pour tous les participants. $SPCH$ = Verbal, MM = Multimodal, NTR = Aucune (contrôle), NVB = Non-verbal. Frd = Amicalité, Hos = Inamicalité, Dom = Dominance.

pour une combinaison de modalités donnée, dans le cas de l'interaction d'une part et des vidéos d'autre part. Une différence significative est seulement observée pour le cas de l'évaluation du modèle de dialogue ($F(1, 46) = 8.59, p = 0.005$), où la différence de reconnaissance est beaucoup plus élevée dans le cas des vidéos par rapport à l'interaction (Q_{Obj} plus élevé de 0.89 en moyenne).

7.3 Discussion

Nous avons réalisé cette évaluation dans plusieurs buts. Nous avons d'abord voulu vérifier que les attitudes étaient globalement reconnues avec le modèle complet ($QR1$). Ensuite, nous avons voulu identifier quelle combinaison de modalités présentait les meilleurs taux de reconnaissance des attitudes par les participants ($QR2$). Enfin, nous avons voulu comparer un mode d'évaluation en interaction avec une évaluation par vidéos ($QR3$).

Nous observons d'abord que la dominance et l'amicalité exprimées par le modèle complet sont globalement bien reconnues ($QR1$). L'inamicalité semble moins bien reconnue (seule la mesure de Q_{Adj2} est significativement différente de la moyenne de l'échelle de mesure), mais les résultats de $QR2$ indiquent que le recruteur exprimant de l'inamicalité avec le modèle complet (cas multimodal) est perçu beaucoup plus hostile que dans les autres cas. Nous concluons donc que notre modèle de recruteur virtuel parvient à exprimer les trois attitudes que nous avons considérées dans cette étude.

Pour $QR2$, nous observons que c'est avec le modèle complet que les attitudes sont les mieux reconnues, pour les trois attitudes considérées. Ces résultats sont en accord avec des études précédentes ayant observé que la combinaison de l'expression verbale et non-verbale permet les meilleurs taux de reconnaissance [Noller, 1985, Bee et al., 2010, Callejas et al., 2014].

De plus, nous observons pour $QR3$ que les scores de reconnaissance sont meilleurs lors de l'évaluation par des vidéos qu'en interaction. Un test supplémentaire a montré que cet effet était particulièrement marqué lorsque l'attitude était uniquement exprimée par le modèle de dialogue. De plus, dans le cas d'un recruteur amical, les participants évaluant des vidéos attribuent au recruteur un plus forte volonté de déstabilisation du candidat. Nous concluons donc que le mode d'évaluation d'un

modèle d'Agent Conversationnel Animé peut influencer les résultats obtenus.

Notre modèle de recruteur virtuel est cependant limité car celui-ci ne modélise pas l'expression d'attitude par le comportement d'écoute et de tour de parole. Or, la manière de réaliser le tour de parole et de produire des *backchannels* influence la perception des attitudes [ter Maat et al., 2010]. Cette limitation est particulièrement importante à prendre en compte quand nous analysons la différence d'évaluation entre le cas en interaction et le cas par vidéos. En effet, le comportement d'écoute n'était pas montré aux participants évaluant des vidéos. Dans le cas où les participants interagissaient avec le recruteur virtuel, le comportement d'écoute neutre a pu atténuer les attitudes ressenties par ces participants, et ainsi contribuer à la différence significative d'évaluation du recruteur entre le cas en interaction et le cas par vidéos. Cependant, la différence d'identification pour le modèle de dialogue seul est tout de même très importante.

CONCLUSION

Lors d'entretiens d'embauche, les recruteurs expriment différentes attitudes. Un système d'entrainement aux entretiens d'embauche doit pouvoir confronter un utilisateur à ces attitudes, en intégrant un recruteur virtuel capable de les exprimer. Dans ce chapitre, nous avons présenté une architecture d'Agent Conversationnel Animé intégrant le modèle de planification de séquences de signaux non-verbaux proposé au chapitre 6 et permettant ainsi le développement d'un recruteur virtuel capable d'exprimer différentes attitudes sociales. Nous avons ensuite implémenté ce recruteur virtuel en intégrant d'autres modèles existants à notre modèle de planification de comportement.

Nous avons réalisé une étude pour valider le modèle proposé en vérifiant que les attitudes exprimées par le recruteur étaient bien reconnues. Les trois attitudes considérées (dominance, amicalité, inamicalité) sont reconnues. Nous avons comparé l'expression de l'attitude avec le modèle de planification de comportement non-verbal seul, avec le modèle de dialogue seul, ou avec les deux. Les résultats montrent que l'expression multimodale des attitudes permet une meilleure identification des attitudes. Enfin, nous avons comparé la perception des attitudes du recruteur par des participants en interaction avec le système et par des participants évaluant des

vidéos. Les résultats montrent que les attitude sont mieux reconnues par les participants évaluant des vidéos, en particulier lorsque les attitudes sont exprimées par le modèle de dialogue uniquement. Ce résultat indique que le mode d'évaluation d'un modèle d'Agent Conversationnel Animé pourrait influer sur les résultats d'une étude. Des études complémentaires sont nécessaires pour vérifier que ce phénomène existe aussi dans d'autres contextes (*ex.* perception des émotions ou de la personnalité d'un ACA).

Synthèse du chapitre

1. Une architecture de recruteur virtuel autonome capable d'exprimer des attitudes lors de sa prise de parole a été présentée. Nous avons implémenté le recruteur en utilisant plusieurs modèles existants de dialogue, de comportement d'écoute, et de tour de parole.
2. Nous avons réalisé une étude pour comparer l'identification des attitudes exprimées avec plusieurs combinaisons de modalités et sous deux modes d'évaluation : en interaction et par vidéos.
   - Le modèle complet - cas multimodal - présente les meilleurs taux d'identification correcte d'attitudes, et permet d'exprimer les attitudes correctement en interaction.
   - Les participants identifient mieux les attitudes en regardant des vidéos qu'en étant en interaction, en particulier dans le cas de l'expression uniquement par le comportement verbal.

8 Une audience virtuelle pour l'entrainement à la prise de parole en public

Dans les chapitres précédents (Chapitres 4 à 7), nous avons étudié l'expression d'attitudes par des Agents Conversationnels Animés. Il est en effet essentiel, dans certaines applications d'entrainement social comme l'entrainement aux entretiens d'embauche, que les Agents Conversationnels Animés puissent confronter l'utilisateur à différentes attitudes. Nous abordons dans ce chapitre la deuxième problématique de notre thèse : la conception et l'évaluation de systèmes d'entrainement de la compétence de prise de parole en public.

Savoir prendre la parole en public est une compétence essentielle dans de nombreuses situations professionnelles et personnelles. Cependant, cette compétence n'est pas innée. Une approche pour s'entrainer à prendre la parole en public consiste à avoir recours à un spécialiste ou à un groupe d'amis ou de connaissances devant lesquels on va pratiquer la prise de parole en public, et qui va pouvoir proposer des conseils d'amélioration après ces séances de pratique [Spence, 2003, Hart et al., 2013]. Cette approche pose plusieurs problèmes. Des amis ou un spécialiste ne sont pas toujours disponibles. De plus, la personnalité et les cultures des différents participants, ou les relations qu'ils entretiennent peuvent influencer le processus d'apprentissage. Enfin, des personnes anxieuses lors de situations de prise de parole en public pourraient être réticentes à s'entrainer devant d'autres personnes. Un système d'audience virtuelle dédié à l'entrainement de la compétence de la prise de parole en public permettrait de pallier ces difficultés, en permettant des sessions d'apprentissage standardisées,

Le travail présenté dans ce chapitre a été réalisé à l'Institute for Creative Technologies de l'Université de Californie du Sud (USC-ICT), sous la supervision de Stefan Scherer et Louis-Philippe Morency. J'ai personnellement réalisé la conception de l'architecture et son implémentation, et mené à bien l'étude, y compris la collecte des vidéos et des données des participants. La collecte des évaluations par des experts (Section 8.3.3.2) et des évaluations objectives du comportement non-verbal des participants (Section 8.3.3.3) a été réalisée par Torsten Wörtwein, stagiaire à l'Institute for Creative Technologies. L'analyse des résultats à été réalisée en collaboration avec les personnes mentionnées ci-dessus.

sans limites de disponibilité, et sans appréhension.

Des systèmes d'audiences virtuelles ont été proposés pour aider des utilisateurs à réduire leur anxiété lors de prises de parole en public [Pertaub et al., 2002, Harris et al., 2002, Grillon et al., 2006]. Cependant, ces audiences virtuelles n'étaient pas interactives. Les membres de l'audience pouvaient exprimer des signaux de différentes natures, mais ceux-ci n'étaient pas liés directement au comportement du participant en train de prendre la parole. De plus, ces systèmes n'ont pas été étudiés dans le but d'améliorer directement la compétence de prise de parole en public des participants, mais afin d'étudier la réduction de l'anxiété liée aux situations de prise de parole en public.

Dans ce chapitre, nous présentons un système d'audience virtuelle interactive, qui permet à un utilisateur de s'entrainer à prendre la parole en public. Les personnages virtuels constituant l'audience fournissent, par leur comportement, des retours en temps réel à l'utilisateur sur sa performance. Dans la prochaine section, nous commençons par présenter les pratiques existantes pour l'amélioration de la compétence de prise de parole en public. Nous proposons ensuite une architecture d'audience virtuelle interactive dédiée à l'amélioration de cette compétence, et indiquons comment nous avons implémenté un système d'audience virtuelle interactive suivant cette architecture. Enfin, nous rapportons les résultats d'une étude où nous avons évalué plusieurs stratégies de retour en temps réel sur la performance de l'utilisateur.

8.1 L'ENTRAINEMENT À LA PRISE DE PAROLE EN PUBLIC

Les compétences sociales comme la prise de parole en public sont des atouts essentiels dans un large nombre de professions et dans la vie quotidienne. La capacité à communiquer dans des situations sociales et publiques peut fortement influencer l'évolution de la carrière d'une personne, aider à construire et maintenir des relations, contribuer à résoudre des conflits, ou encore emporter l'avantage lors de négociations. La maitrise de son comportement non-verbal et de sa voix, c'est à dire des gestes, postures, expressions faciales, ou encore de sa prosodie, sont des éléments clés de la réussite de prises de parole en public ou de communications interpersonnelles [Batrinca et al., 2013]. Par exemple, l'analyse de la communication non-verbale d'un médecin permet de prédire la satisfaction de ses patients [DiMatteo et al., 1986], et

l'analyse de la communication non-verbale d'un négociateur permet de prédire la réussite de ses négociations [Park et al., 2012].

Cependant, la maitrise de sa propre communication non-verbale lors de prises de parole en public et autres situations sociales n'est pas une compétence innée pour tous, et elle peut être améliorée en s'entrainant régulièrement [Hart et al., 2013]. De plus, certaines situations sociales, la prise de parole en public en particulier, peuvent générer de l'anxiété chez certaines personnes, rendant cette expérience à la fois difficile et désagréable. Une manière de s'entrainer consiste à faire des présentations dans des environnements familiers (*ex.* devant des membres de sa famille, amis, collègues, *etc.*), et à recevoir des conseils après la présentation. Les personnes constituant l'audience peuvent aussi produire un retour indirect pendant des présentations par leur comportement non-verbal [MacIntyre et al., 1997]. Par exemple, une audience attentive et engagée devant une présentation pourra montrer des signes d'attention (*ex.* postures dirigées vers le présentateur, hochements de tête, *etc.*), ou inversement pourra montrer des signes de désintérêt, voire même de désaccord (*ex.* regards distrait, hochements de tête horizontaux). D'autres pratiques pour améliorer ses compétences sociales consistent à apprendre par le biais de livres spécialisés, à s'entrainer dans des séances de jeu de rôle avec des spécialistes, et à enregistrer des vidéos de ses performances pour pouvoir les critiquer *a posteriori* [Spence, 2003, Peterson, 2005, Hart et al., 2013].

Comme nous l'avons détaillé dans la section 3.1, des systèmes utilisant des Agents Conversationnels Animés ont été proposés pour l'entrainement des compétences sociales [Anderson et al., 2013, Talbot et al., 2012, Lane et al., 2013, Hoque et al., 2013, Swartout et al., 2013]. De tels systèmes se sont révélés être des outils favorisant un bon engagement des apprenants, or l'engagement est un facteur influençant favorablement le processus d'apprentissage [Johnson et al., 2000, Rowe et al., 2010, Hart et al., 2013]. De plus, le comportement d'Agents Conversationnels Animés peut être précisément contrôlé lors de différentes séances d'entrainement, ce qui permet le développement d'approches d'apprentissage standardisées pour l'entrainement de compétences sociales. Cependant, il n'existe pas aujourd'hui de système dédié à l'amélioration de compétence de prise de parole en public.

Dans la section suivante, nous présentons une architecture de système d'audience virtuelle interactive pour l'amélioration de la compétence de prise de parole en pu-

blic.

8.2 Architecture et implémentation d'une audience virtuelle interactive

Lors de la conception d'une architecture de système pour l'entrainement à la prise de parole en public, nous nous sommes donné plusieurs objectifs. Tout d'abord, le système devait pouvoir déterminer la *performance multimodale* de l'utilisateur lors d'une présentation. Comme il a été démontré dans [Batrinca et al., 2013], la performance globale d'un présentateur est corrélée avec de nombreux éléments du comportement non-verbal et verbal (*ex.* regards dirigés vers l'audience, intonation claire, utilisation de gestes, *etc.*), qui peuvent être analysés de manière automatique. Ensuite, le système devait pouvoir fournir un *retour en temps réel* à l'utilisateur sur sa performance. Enfin, nous voulions pouvoir, dans le futur, explorer l'influence de différentes tailles et configurations d'audience (*ex.* une audience répartie sur plusieurs écrans). Nous voulions aussi pouvoir réaliser des expériences utilisant soit une détection de comportement automatique, par le biais d'un assortiment de senseurs hétérogènes distribués sur plusieurs machines, soit un *magicien d'Oz* (*i.e.* une personne contrôlant partiellement le système).

L'architecture que nous proposons est représentée sur la Figure 8.1. Un ou plusieurs logiciels de perception multimodale permettent de déterminer des informations *bas-niveau* sur le comportement du présentateur à des intervalles réguliers (*ex.* angle de la tête, détection d'un geste, fréquence fondamentale de la voix). Alternativement, un *magicien d'Oz* peut se charger de détecter les signaux non-verbaux du présentateur. Ces données sont envoyées dans le format de perception multimodale PML [Scherer et al., 2012b] (*P*erception *M*arkup *L*anguage) à un module central, l'*agrégateur*, qui en dérive des variables permettant de caractériser la performance de prise de parole de l'utilisateur (*ex.* quantité de regards dirigés vers l'audience, quantité d'hésitations dans le discours). Par exemple, Batrinca *et al.* ont démontré qu'un regard dirigé vers l'audience est corrélé avec une bonne performance de prise de parole en public [Batrinca et al., 2013]. Une variable pouvant caractériser la qualité du comportement de regard de l'utilisateur pourra alors être la proportion de regards dirigés vers l'audience, dans une fenêtre de temps donnée (*ex.* les dix dernières secondes). Ces variables, que nous appelons *descripteurs haut-niveau*, sont

définies dans l'intervalle [0; 1], une valeur de 0 correspondant à une performance très mauvaise, tandis que 1 indique une très bonne performance pour le comportement considéré. Ces descripteurs peuvent être affichés directement par des éléments de retour direct (*ex.* une barre remplie si la performance est bonne, vide sinon). Le but de ces éléments de retour direct est de donner au participant une indication claire et objective sur sa performance. Par exemple, une barre de couleur pourra afficher la proportion exacte des regards que le participant a dirigés vers l'audience.

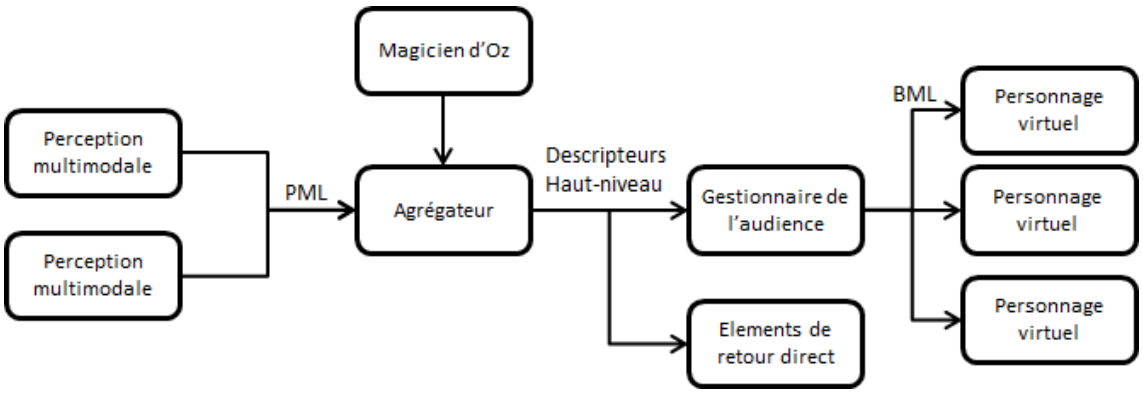


FIGURE 8.1: Architecture du système d'audience virtuelle. L'architecture distribuée que nous avons adoptée permet d'utiliser plusieurs logiciels de perception multimodale simultanément (*ex.* un logiciel dédié à l'analyse du flux vidéo d'une Webcam, un autre logiciel dédié á l'analyse du flux audio d'un microphone), et de distribuer les personnages virtuels sur plusieurs écrans.

Les *descripteurs haut-niveau* sont aussi envoyés à un module déterminant le comportement des différents membres de l'audience, le *gestionnaire d'audience*. Chaque membre de l'audience est associé à un profil de comportement, qui détermine, en fonction de l'évolution de la performance de l'utilisateur, quand ce personnage va afficher des signaux non-verbaux. Ces profils définissent les signaux non-verbaux que les personnages affichent lorsque certaines conditions sont remplies (*ex.* sourire lorsque le participant regarde dans la direction du personnage). Plus précisément, un profil consiste en un ensemble de 3-uplets sous le format $< Descripteur; Signal; Conditions >$. *Descripteur* est l'identifiant du descripteur haut-niveau considéré, *Signal* est un message BML décrivant les signaux non-verbaux déclenchés, et *Conditions* décrivant les conditions dans lesquelles *Signal* est déclenché. Ci-dessous, nous présentons un exemple de profil de comportement. La première ligne indique que lorsque le descripteur de la quantité d'hésitations (`q_hesitations`) est compris dans l'intervalle $[0.8, 1.0]$, le personnage doit sourire (`smile`). La deuxième ligne indique que lorsque

le descripteur de la quantité de regards dirigés vers l'audience (`q_regard`) est compris dans l'intervalle $[0.0, 0.5]$, le personnage doit hocher la tête horizontalement (`shake`).

```
<q_hesitations,<face type="smile"/>,in(0.8,1.0)>
<q_regard,<head type="shake"/>,in(0.0,0.5)>
```

Ainsi, les personnages virtuels constituant l'audience peuvent être utilisés pour fournir, par leur comportement, une indication indirecte au participant sur sa performance. Dans notre implémentation, les membres de l'audience peuvent changer de posture (se pencher vers le présentateur en se montrant attentif, se tenir droit sur son siège, ou bien adopter une posture relâchée), mais aussi régulièrement hocher la tête verticalement ou horizontalement. Les personnages peuvent aussi signaler à l'utilisateur qu'il ne les a pas regardés depuis trop longtemps en s'éclaircissant la voix (animation et son). Les comportements positifs (se rapprocher de l'utilisateur est un signe d'amicalité, hocher la tête est un signe d'accord, voir Section 2.2.3) sont déclenchés lorsque la performance du présentateur est bonne, tandis que les comportements négatifs (se pencher en arrière est un signe d'inamicalité, hocher la tête horizontalement est un signe de désaccord, voir Section 2.2.3) se déclenchent lorsque la performance du présentateur est mauvaise. Afin que les personnages n'envoient pas tous les mêmes signaux simultanément, nous avons défini des valeurs de déclenchement des signaux non-verbaux différentes pour les membres de l'audience. Nous avons réparti ces valeurs de manière uniforme dans l'intervalle d'évolution des descripteurs haut-niveau ($[0; 1]$) afin que le ratio de membres de l'audience affichant des signaux positifs par rapport à ceux affichant des signaux négatifs soit directement lié à la performance de la personne.

Nous avons implémenté notre système à partir de la plate-forme Virtual Human Toolkit de l'Institute for Creative Technologies[1]. Le logiciel de perception multimodale que nous avons utilisé est Multisense [DeVault et al., 2014]. Celui-ci intègre plusieurs composants dédiés au suivi de la tête et du regard, de la détection des expressions faciales, et de l'analyse d'un signal audio. La scène 3D d'audience virtuelle est affichée par le moteur de jeu Unity, et les personnages virtuels sont animés par la plateforme d'animation Smartbody [Shapiro, 2011], compatible avec le format BML. La configuration du système repose sur un ensemble de fichiers permettant de

[1] `https://vhtoolkit.ict.usc.edu/`

FIGURE 8.2: Capture d'écran de l'audience virtuelle. En haut, une barre de couleur indique la performance de l'utilisateur (Si la barre est totalement verte, la performance est très bonne. Si elle est totalement rouge, celle-ci est très mauvaise.). Les personnages virtuels changent de posture en fonction de la performance de l'utilisateur.

définir les positions et apparences des différents membres de l'audience, leurs profils de comportements, les éléments de retour direct, et les règles de calcul des descripteurs haut-niveau. Cela permet de modifier la configuration du système aisément même lorsque celui-ci est en cours d'exécution. De plus, les différents composants du système peuvent être répartis sur plusieurs ordinateurs (*ex.* dédier un ordinateur au traitement vidéo, un ordinateur au traitement audio et un dernier à l'audience virtuelle). La scène 3D consiste en une pièce simple arrangée sur trois niveaux, permettant d'organiser l'audience sur plusieurs lignes. Dans notre implémentation, l'audience était constituée de quinze personnages, utilisant tous le même modèle de personnage mais avec une combinaison de vêtements et de couleur de cheveux choisies aléatoirement. L'audience était répartie sur deux écrans adjacents. La figure 8.2 présente une capture d'écran d'une des deux parties de l'audience.

Nous avons réalisé une étude pour évaluer notre système sous trois perspectives différentes : celle des apprenants eux-mêmes, celle d'experts en prise de parole en public,

et une perspective objective d'évaluation par des mesures objectives du comportement des participants. Nous présentons cette étude dans la section suivante.

8.3 ÉVALUATION DU SYSTÈME

Un facteur qui peut améliorer le succès d'un processus d'apprentissage est l'engagement de l'utilisateur [Moreno et al., 2001, Rowe et al., 2010, Hart et al., 2013]. Nous avons ainsi voulu mesurer l'engagement des participants s'entrainant avec l'audience virtuelle, ainsi que leur perception de la difficulté d'utilisation du système. En effet, un système trop difficile à utiliser pourrait démotiver les utilisateurs. Afin de pouvoir juger les différents aspects de la performance des participants, nous avons eu recours à des experts en prise de parole en public. Certains concepts permettant de caractériser et d'évaluer la qualité d'une prise de parole sont complexes et difficiles à formaliser (*ex.* la *confiance en soi* que dégage le présentateur). Nous avons ainsi interrogé des experts pour évaluer ces aspects complexes de la performance des participants. Enfin, nous avons complété notre étude par l'analyse de mesures objectives du comportement des présentateurs. Pour cela, nous avons sélectionné des comportements simples à détecter et qui ont été identifiés par des travaux précédents comme des facteurs importants d'une bonne performance de prise de parole en public [Batrinca et al., 2013] : la quantité de regards adressés à l'audience, et la quantité d'hésitations dans le discours (*pause fillers*). Nous précisons nos trois perspectives d'évaluation du système en posant les questions suivantes :

Q1 : Du point de vue de l'utilisateur, à quel point l'entrainement à la prise de parole avec une audience virtuelle est une expérience engageante, et à quel point cet exercice est-il difficile ?

Q2 : Selon des experts en prise de parole en public, les participants améliorent-ils leurs compétences en prise de parole après s'être entrainés avec une audience virtuelle ?

Q3 : L'audience virtuelle permet-elle aux participants d'améliorer certains aspects de leur comportement, mesurés objectivement, en particulier la quantité d'hésitations et de regards dirigés en direction de l'audience ?

Nous avons voulu analyser à quel point une audience interactive permet d'améliorer l'expérience d'apprentissage et l'évaluation d'un système d'entrainement par les

utilisateurs, en comparaison à une audience non réactive ou utilisant des éléments de retour génériques. Nous avons défini trois conditions correspondant à trois stratégies de retour à l'utilisateur : (1 - contrôle) pas de comportement de la part de l'audience, pas de retour sur la performance du participant, (2 - retour direct) retour visuel direct sur la performance du participant par le biais de jauges de couleur, et (3 - audience interactive) retour indirect par le comportement non-verbal de l'audience virtuelle. Une autre condition envisageable aurait été de réunir une audience réelle ou des experts qui auraient pu fournir des retours adaptés aux participants. Cependant, cette option n'était pas réalisable pour des raisons logistiques.

Pour notre étude, nous avons concentré spécifiquement l'entrainement des participants sur deux aspects de leur comportement : la quantité d'hésitations et le comportement de regard. Nous avons choisi ces deux aspects du comportement sur les conseils d'experts de l'association Toastmasters[2], dont le but est d'entrainer les personnes à la prise de parole en public, et car ils ont été identifiés dans des travaux précédents comme corrélés avec la performance d'une présentation [Batrinca et al., 2013]. De plus, ces deux aspects du comportement sont faciles à définir, à identifier et à quantifier, permettant ainsi une évaluation objective de la performance des participants sur ces deux aspects.

8.3.1 Protocole expérimental

Afin d'étudier l'impact de ces trois stratégies de retour, nous avons organisé une étude suivant une méthode *pre-test/post-test*, c'est à dire que nous mesurons l'amélioration de la performance de prise de parole des participants entre une évaluation préliminaire (*pre-test*) et une évaluation ultérieure (*post-test*), entrecoupées de séances d'entrainement. Deux séances d'entrainement sont réalisées, la première dédiée à réduire la quantité d'hésitations dans le discours, la seconde dédiée à l'amélioration du comportement de regard. L'organisation de l'étude est présentée sur la figure 8.3. Les participants sont répartis aléatoirement entre trois conditions, correspondant à trois configurations du système implémentant trois différentes stratégies de retour lors des présentations d'entrainement.

[2] www.toastmasters.org

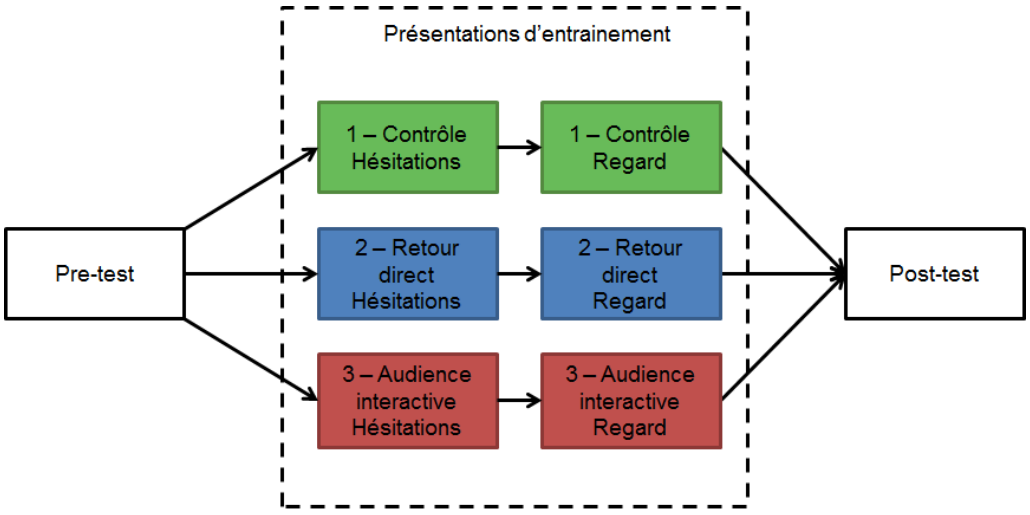


FIGURE 8.3: Protocole expérimental. Dans chaque condition, les participants sont d'abord évalués, puis s'entrainent deux fois (un entrainement centré sur le comportement de regard et un autre sur les hésitations), puis sont à nouveau évalués.

8.3.1.1 Configurations du système

Lors de l'évaluation, le système pouvait être configuré dans les trois modes suivants :

1. **Contrôle (Audience virtuelle non-interactive) :** Pas de retour du système : pas d'indications visuelles directes et pas de comportement de la part de l'audience (seulement une animation neutre et des clignements des yeux).

FIGURE 8.4: Audience virtuelle dans la condition de contrôle.

2. **Retour direct (Jauges de performance) :** Un retour direct sur la performance du participant (par rapport au comportement entrainé : regard ou

hésitations) est réalisé en temps réel sous la forme d'une jauge colorée au-dessus de l'audience virtuelle. Pas de comportement de la part de l'audience (seulement une animation neutre et des clignements des yeux).

FIGURE 8.5: Audience virtuelle dans la condition de retour direct. Par rapport à la condition de contrôle, une jauge colorée est ajoutée au-dessus de l'audience virtuelle.

3. **Retour non-verbal (Audience virtuelle interactive) :** Un retour sur la performance du participant (par rapport au comportement entrainé : regard ou hésitations) est réalisé en temps réel sous la forme de signaux non-verbaux de la part de l'audience. L'audience montre des comportements positifs lorsque la performance du participant est bonne (hochements de tête verticaux, postures vers l'avant), et des comportements négatifs lorsque la performance du participant est mauvaise (posture relâchée, hochements de tête horizontaux, éclaircissements de gorge).

FIGURE 8.6: Audience virtuelle dans la condition de retour non-verbal. Selon la performance de l'utilisateur, les personnages vont adapter leur posture (se pencher en avant ou en arrière) et hocher la tête verticalement ou horizontalement.

Pour les configurations (2 - Retour direct) et (3 - Retour non-verbal), afin de s'assurer d'une bonne reconnaissance, un *magicien d'Oz* était en charge de détecter le comportement du présentateur (direction de regard ou hésitations). En effet, bien que des travaux aient été alors entamés pour intégrer les composants logiciels permettant de reconnaitre automatiquement ces comportements, ceux-ci n'étaient pas encore achevés.

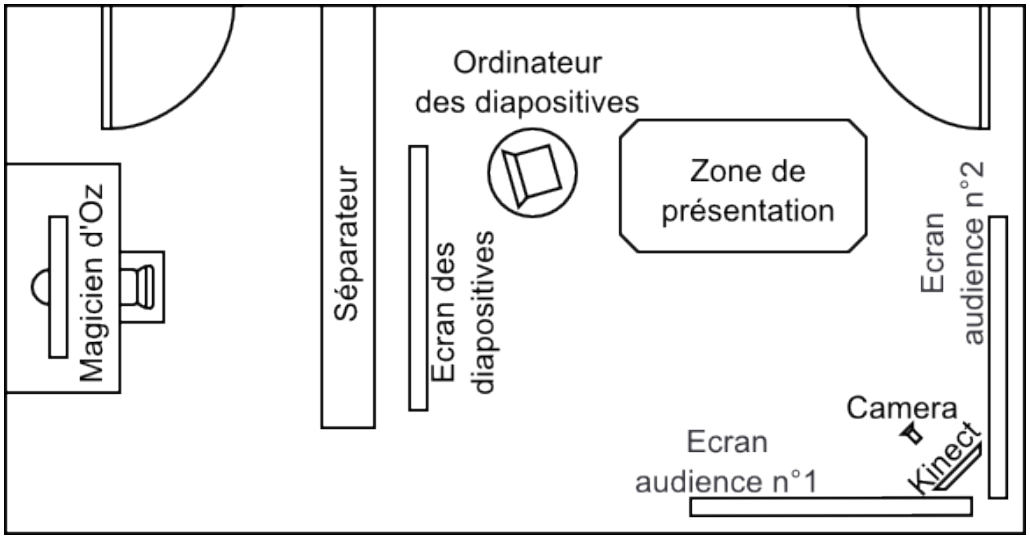


FIGURE 8.7: Organisation de la salle d'étude. L'écran des diapositives est disposé à droite des participants, tandis qu'une partie de l'audience est en face d'eux, et une autre partie à leur gauche. Cette disposition force les participants à devoir déplacer leur tête lors des présentations, ce qui permet de créer une situation d'apprentissage propice à l'entrainement du comportement de regard.

L'audience virtuelle était affichée à l'aide de deux projecteurs afin de produire une audience à taille réelle. Les participants devaient se tenir dans une zone de présentation marquée au sol. La disposition de cette zone et des différents écrans (écran des diapositives et des deux audiences) a été réalisée de manière à ce que les participants soient forcés de devoir déplacer leur tête entre les extrémités de l'audience et entre les diapositives, afin de rendre la tâche de reconnaissance de direction du regard plus simple au *magicien d'Oz*. Un écran et un projecteur étaient dédiés à l'affichage des diapositives de la présentation du participant. Un ordinateur contrôlant les diapositives était disposé à côté de la zone de présentation des participants, et ceux-ci disposaient d'une télécommande sans fil leur permettant de passer à la prochaine diapositive. Les participants étaient enregistrés avec un microphone serre-tête, une caméra dirigée vers le visage des participants, et une Microsoft Kinect placée au

niveau du milieu de l'audience orientée vers les participants. La figure 8.7 présente l'organisation de la salle d'étude et l'emplacement des différents capteurs.

8.3.1.2 Déroulement de l'expérience

Quelques jours avant leur participation à l'étude, des instructions étaient envoyées aux participants leur indiquant qu'ils devraient réaliser plusieurs présentations de cinq minutes sur deux sujets différents. Le premier sujet consistait à présenter la ville et la région de Los Angeles (la ville où l'étude a été réalisée), le deuxième sujet consistait à présenter un produit de beauté. Des fichiers comprenant cinq diapositives sur ces deux sujets étaient fournies ainsi qu'une proposition d'un résumé de présentation (les résumés sont disponibles en annexe B) afin de permettre aux participants de se préparer. Cependant les participants avaient la liberté d'improviser leur texte et le détail de leur présentation, la seule contrainte étant de parler du sujet concerné.

Le jour de l'évaluation, la première étape pour les participants était de remplir un ensemble de questionnaires d'auto-évaluation de leur compétence de prise de parole en public (présentés dans la section 8.3.3 et reproduits dans l'annexe B). Ensuite, chaque participant donnait quatre présentations (voir Figure 8.3). La première et la quatrième présentation sont les présentations *pre-test* et *post-test*. Dans ces deux présentations, les participants présentaient le premier sujet en face de l'audience virtuelle configurée dans le mode de contrôle. Entre ces deux tests, les participants faisaient deux présentations sur le deuxième sujet, avec pour but d'améliorer leur comportement de regard et leur quantité d'hésitations dans le discours. Indépendamment de la condition dans laquelle se trouvait le participant, une feuille d'information lui était fournie, contenant les conseils des experts de Toastmasters par rapport à l'aspect du comportement considéré (*i.e.* soit le regard, soit les hésitations. Ces documents sont reproduits en annexe B). Le système, en revanche, était configuré en fonction de la condition assignée au participant : un tiers des participants s'entrainaient donc avec le système configuré dans le mode de contrôle, un tiers avec le mode de retour direct, et le dernier tiers avec l'audience interactive. Les hésitations produites par les participants et la direction de leur regards étaient détectés par un *magicien d'Oz*.

Après la dernière présentation (*post-test*), les participants devaient remplir des questionnaires afin d'évaluer leur impression du système (voir Section 8.3.3). Les impressions des participants étaient ensuite recueillies de manière informelle, puis les participants étaient payés 25$ pour leur participation, et remerciés. En moyenne, l'expérience prenait une cinquantaine de minutes.

8.3.2 Participants et données collectées

Les participants étaient recrutés par le biais de Craigslist[3], dans la région de Los Angeles. Au total, 47 participants ont pris part à l'expérience (29 hommes et 18 femmes), d'un âge moyen de 37 ans (σ = 12.05). Parmi les 47 participants, 30 ont rapporté avoir été à l'université. Des problèmes techniques lors de deux enregistrements nous ont empêché d'utiliser les données de deux participants. Nous disposons donc d'enregistrements pour un nombre total de 45 participants, avec 15 participants assignés à la condition de contrôle (audience non interactive, pas de retour direct), 14 à la condition de retour direct, et 16 à l'audience virtuelle interactive. En moyenne, les présentations de *pre-test* duraient 3 minutes et 57 secondes (σ=1 minute et 56 secondes) et les présentations de *post-test* en moyenne 3 minutes et 54 secondes (σ=2 minutes et 17 secondes). Nous n'avons pas observé de différence significative entre ces deux durées.

8.3.3 Mesures

8.3.3.1 Questionnaires d'auto-évaluation

Les participants complétaient plusieurs questionnaires avant la présentation *pre-test* : un questionnaire démographique ; le questionnaire de *« Personal Report of Confidence as a Speaker (PRCS) »* [McCroskey, 1970], et le questionnaire *« Self-Statements During Public Speaking »* [Hofmann & DiBartolo, 2000], utilisés pour évaluer la confiance en soi et l'anxiété du participant par rapport aux situations de prise de parole en public en général [Hook et al., 2008] ; un questionnaire court pour estimer la personnalité des participants selon le modèle du Big-5 (*« Big Five Inventory »*) [Rammstedt & John, 2007]. Après la dernière présentation, les participants

[3] www.craigslist.org

remplissaient le questionnaire *« Positive and Negative Affect Schedule (PANAS) »* servant à évaluer les émotions ressenties par le participant [Thompson, 2007], une version adaptée d'un questionnaire de mesure d'immersion [Jennett et al., 2008] et un questionnaire d'auto-évaluation réalisé pour l'expérience dont l'objet était de recueillir les impressions des participants concernant le système. Tous ces documents sont disponibles en annexe B.

8.3.3.2 Évaluation par des experts

Afin de comparer les présentations *pre-test* et *post-test*, trois experts de l'assocation Toastmasters ont été invités à juger les présentations des participants et rémunérés 125$ pour leur participation. Leur âge moyen était de $43,3$ ans ($\sigma = 11,5$), un des experts était une femme et deux des hommes. Les experts ont évalué leur propre expérience et leur propre confiance en eux lors de présentations en public sur des échelles de Likert à 7 points. En moyenne, ils ont rapporté être très à l'aise lors de présentations en public ($\mu = 6.3$, avec 1 - *not comfortable*, 7 - *totally comfortable*). Ils rapportent avoir beaucoup d'expérience sur le sujet ($\mu = 6$, avec 1 - *no experience*, 7 - *a lot of experience*). Chaque expert a déclaré avoir donné plus de onze présentations devant une audience au cours des deux dernières années.

Les experts avaient la tâche d'évaluer si l'entrainement avait permis aux participants d'améliorer leurs compétences de prise de parole en public. Lors de cette évaluation, ils ne connaissaient pas la condition d'entrainement dans laquelle les participants se trouvaient. Les vidéos des présentations *pre-test* et *post-test* de chaque participant étaient présentées côte à côte pour une comparaison directe. L'ordre de présentation des vidéos (vidéo *pre-test* à gauche de la vidéo *post-test* ou l'inverse), ainsi que l'ordre des participants étaient cependant aléatoire. Chaque vidéo était cadrée sur le haut du corps afin de bien voir les expressions faciales et les gestes, et comportait de plus une vignette avec une vision globale de la pièce pour analyser les postures, les déplacements et l'utilisation de la pièce par le participant (voir Figure 8.8).

Les experts ont évalué différents aspects des performances des participants sur des échelles de Likert à sept points, identifiés précédemment comme des aspects importants du comportement lors de prises de parole réussies [Schreiber et al., 2012, Batrinca et al., 2013, Scherer et al., 2012a, Rosenberg & Hirschberg, 2005] :

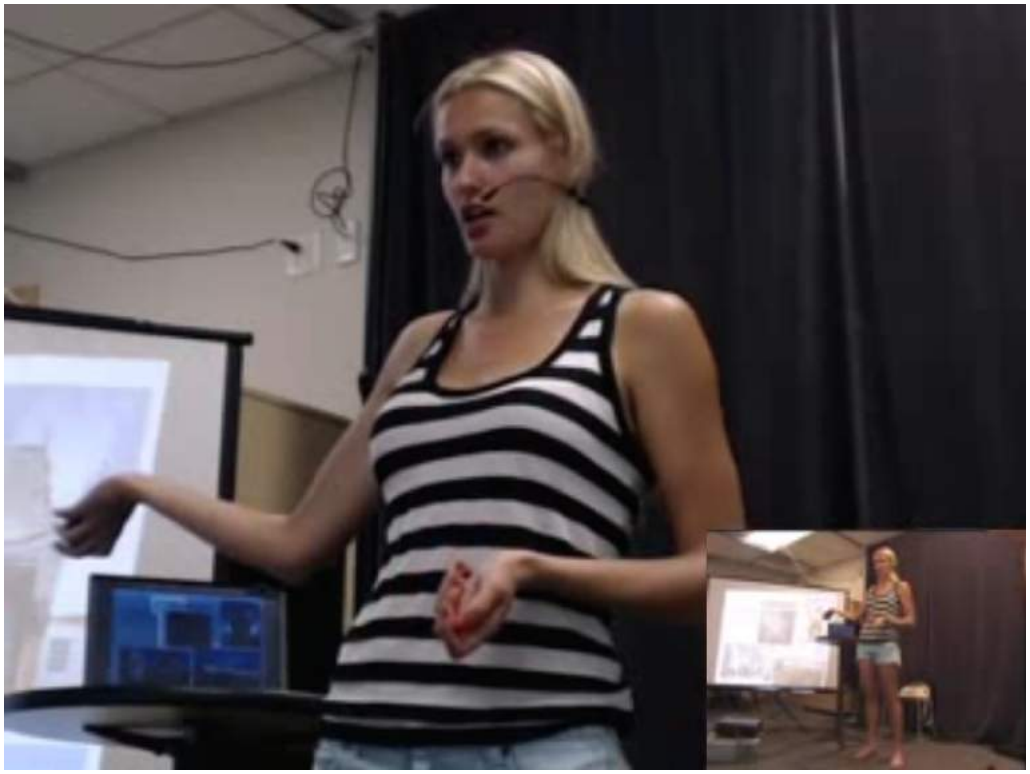


FIGURE 8.8: Exemple de vidéo vue par un expert.

1. Regard (*Eye contact*)
2. Posture (*Body posture*)
3. Débit de parole (*Flow of Speech*)
4. Gestes (*Gesture usage*)
5. Intonation (*Intonation*)
6. Confiance en soi (*Confidence Level*)
7. Utilisation de la pièce (*Stage Usage*)
8. Hésitations (*Avoid Pause Fillers*)
9. Structure de la présentation (*Presentation Structure*)
10. Performance globale (*Overall Performance*)

Les échelles de Likert étaient utilisées ainsi : pour l'aspect du comportement considéré, une valeur de 1 indique que la personne est meilleure sur la vidéo de gauche,

une valeur de 7 indique que la personne est meilleure sur la vidéo de droite, tandis qu'une valeur de 4 indique que la personne se comporte aussi bien pour l'aspect considéré entre les deux vidéos. Après avoir collecté ces données, nous les avons transformé pour annuler l'effet de l'ordre aléatoire de présentation des vidéos. Une valeur de -3 indique donc, pour le comportement considéré, que la personne se comporte bien mieux dans la présentation *pre-test* (*i.e.* avant l'entrainement), une valeur de 0 indique un comportement équivalent entre le *pre-test* et le *post-test*, et une valeur de 3 indique que cet aspect est bien meilleur dans la présentation *post-test*.

8.3.3.3 Mesures objectives

Pour compléter les auto-évaluations des participants par des questionnaires (Section 8.3.3.1) et les jugements par des experts (Section 8.3.3.2), nous avons analysé les performances de prise de parole des participants par le biais de deux mesures objectives, correspondant aux aspects du comportement entrainés pendant l'étude : la quantité de regards dirigés vers l'audience, et la quantité d'hésitations dans le discours. Avec l'outil Elan [Wittenburg et al., 2006], deux annotateurs ont manuellement segmenté les périodes des vidéos *pre-test* et *post-test* de chaque participant où leurs regards étaient dirigés vers l'audience (*resp.* vers une autre direction), et les périodes où la personne hésite en parlant (*resp.* ne parle pas ou parle sans hésiter). Pour les deux aspects, nous observons un bon accord inter-annotateur mesuré en calculant l'α de Krippendorf sur un sous-ensemble de quatre vidéos annotées indépendamment par les deux annotateurs : respectivement, $\alpha = 0.751$ pour les regards dirigés vers l'audience, et $\alpha = 0.957$ pour les hésitations.

A partir de ces annotations manuelles, nous calculons le ratio $r_{eye} \in [0, 1]$ entre les instants où le participant regarde vers l'audience et où il regarde ailleurs, avec $r_{eye} = 0$ si le participant ne regarde jamais l'audience et $r_{eye} = 1$ s'il la regarde en permanence. La quantité d'hésitations $q_{pausefillers}$ est quand à elle normalisée par rapport à la durée totale de la présentation. L'amélioration mesurée objectivement pour ces deux comportements est mesurée par la différence normalisée nd entre les présentations *pre-test* et *post-test* que l'on pose de la manière suivante :

$$nd = \frac{post - pre}{post + pre} \in [-1, 1]$$

8.4 RÉSULTATS

Dans cette section, nous rapportons les résultats de notre évaluation en réponse aux trois questions de recherche que nous avons introduit dans la section 8.3, et faisant référence respectivement à l'auto-évaluation des participants (**Q1**), aux jugements d'experts (**Q2**) et aux mesures objectives d'amélioration (**Q3**).

8.4.1 Q1 - Auto-évaluation

Nous rapportons ici les résultats significatifs obtenus en analysant les réponses des participants aux questionnaires, d'abord en analysant les différences observées entre les différentes conditions, puis indépendamment des conditions. Les questions des questionnaires sont identifiées par le symbole SA_{Qi}, où i est l'indice de la question. Par souci de concision, nous ne rapportons que les résultats significatifs. Les questionnaires et les résultats sont détaillés en annexe B (pages 219-222).

Différences entre conditions : Pour chacune des 31 questions d'auto-évaluation, nous réalisons une analyse de la variance à un facteur (*one-way ANOVA*) et des T-tests de Student bilatéraux entre les trois conditions et rapportons les résultats significatifs.

Nous observons une différence significative entre les trois conditions pour la question SA_{Q1}, à savoir à quel point l'audience a retenu l'attention des participants (SA_{Q1} ; $F(2, 44) = 4.628$, $p = 0.015$). Les participants ayant utilisé l'audience interactive ($\mu = 4.50$, $\sigma = 0.52$) ont senti qu'elle retenait leur attention significativement plus que les participants dans la condition de contrôle ($\mu = 3.44$, $\sigma = 1.09$; $t(30) = 0.86$, $p = 0.001$) et dans la condition de retour direct ($\mu = 3.60$, $\sigma = 1.40$; $t(29) = 1.04$, $p = 0.023$). Nous ne trouvons pas de différence significative entre la condition de contrôle et celle de retour direct.

Nous n'observons pas de différence significative pour la question (SA_{Q6} ; $F(2, 44) = 2.229$, $p = 0.120$), à savoir à quel point les participants étaient conscients lors de la présentation qu'ils présentaient devant une audience virtuelle. Cependant, des T-test de Student révèlent que les participants ayant utilisé l'audience interactive ($\mu = 3.56$, $\sigma = 1.63$) étaient plus conscients de présenter devant une audience virtuelle que dans la condition de contrôle ($\mu = 2.62$, $\sigma = 1.20$; $t(30) = 1.43$, $p = 0.049$).

Enfin, nous n'observons pas de différence significative pour la question SA_{Q17} (SA_{Q17} ; $F(2, 44) = 2.745$, $p = 0.075$), évaluant à quel point les participants ont trouvé l'expérience stimulante (*challenging*). Mais des T-tests de Student révèlent une différence entre les participants ayant utilisé l'audience interactive ($\mu = 2.94$, $\sigma = 1.12$) et ceux ayant utilisé l'audience passive dans la condition de contrôle ($\mu = 2.00$, $\sigma = 1.03$; $t(28) = 1.02$, $p = 0.025$).

Résultats indépendants des conditions : Indépendamment des trois conditions d'apprentissage, nous étudions les réponses des participants aux questionnaires afin d'évaluer leurs impressions générales et s'ils expriment de l'intérêt envers l'utilisation d'une plate-forme d'audience virtuelle pour l'entrainement à la prise de parole en public. Pour cela, nous réalisons des tests T de Student bilatéraux, contre l'hypothèse nulle que la réponse moyenne des participants aux questions coïncide avec la moyenne de l'échelle de réponse (c'est à dire $\mu = 3.00$).

Lorsqu'il leur était demandé à quel point ils étaient concentrés sur l'audience virtuelle, les participants ont répondu significativement au dessus de la moyenne (SA_{Q2} ; $\mu = 3.87$, $\sigma = 1.19$; $t(46) = 1.19$, $p < 0.001$). Nous observons la même chose pour la question SA_{Q19}, qui demandait si les participants trouvaient facile de prendre la parole devant l'audience virtuelle ($\mu = 4.02$, $\sigma = 0.99$; $t(46) = 0.99$, $p = 0.000$).

Les réponses des participants sont significativement plus fortes que la moyenne à la question leur demandant s'ils souhaiteraient répéter cette expérience (SA_{Q28} ; $\mu = 4.64$, $\sigma = 0.79$; $t(46) = 0.79$, $p < 0.001$). Les participants ont jugé que le système d'audience virtuelle était très utile pour s'entrainer à la prise de parole en public (SA_{Q29} ; $\mu = 4.81$, $\sigma = 0.50$) ; $t(46) = 0.50$, $p < 0.001$) et que si un tel outil était disponible, ils l'utiliseraient pour améliorer leurs compétences en prise de parole en public (SA_{Q30} ; $\mu = 4.77$, $\sigma = 0.56$; $t(46) = 0.56$, $p = 0.000$).

Enfin, les participants ont noté l'immersion qu'ils ressentaient lors de l'expérience significativement au-dessus de la moyenne (SA_{Q31} ; $\mu = 7.61 > 5$, $\sigma = 2.49$; $t(43) = 2.49$, $p < 0.001$).

8.4.2 Q2 - Évaluations par les experts

Dans cette section, nous rapportons les différences observées dans les jugements produits par les experts entre les vidéos tournées avant l'entrainement avec le système

d'audience virtuelle (*pre-test*) et celles tournées après (*post-test*), en fonction des trois conditions. Pour chacun des dix aspects étudiés d'une bonne prise de parole en public (voir Section 8.3.3.2), nous réalisons une analyse de la variance à un facteur (*one-way ANOVA*) et des tests T de Student bilatéraux entre les différentes conditions.

Comme nous l'avons indiqué en section 8.3.3.2, les mesures présentées sont dans l'intervalle $[-3, 3]$, une valeur négative indique une diminution de la performance entre la présentation *pre-test* et *post-test*, tandis qu'une valeur positive indique une amélioration. Nous ne rapportons les résultats des ANOVA et tests T de Student que lorsque ceux-ci se sont révélés significatifs, mais la table 8.1 et la figure 8.9 présentent les valeurs moyennes et écart types des dix aspects étudiés.

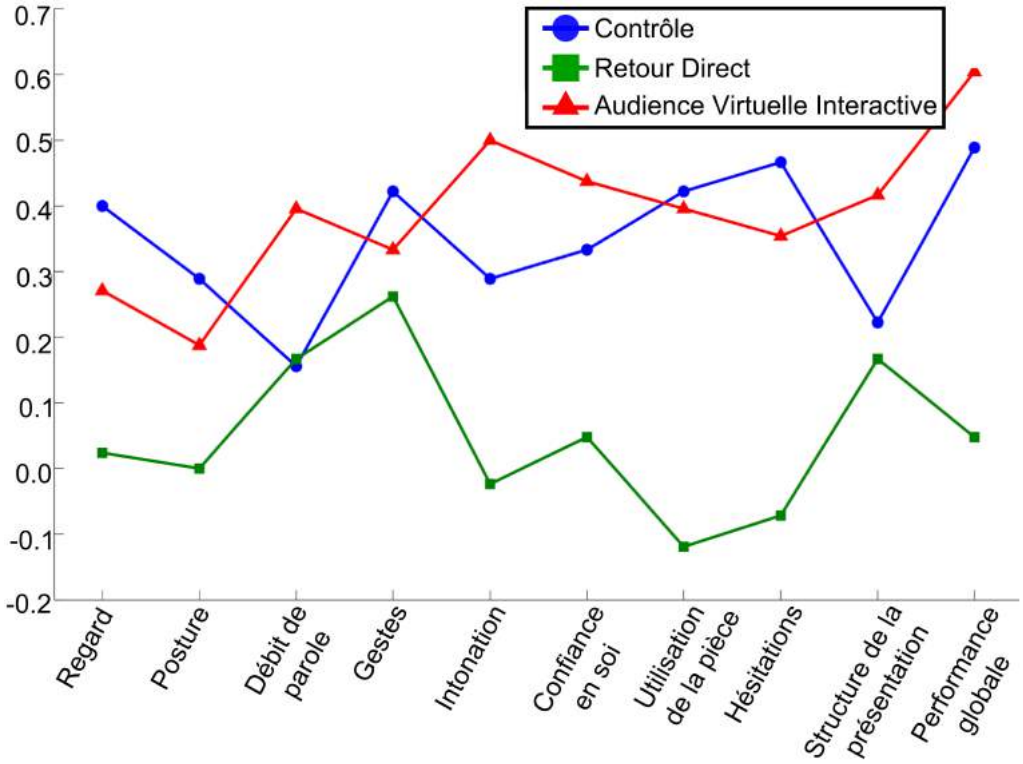


FIGURE 8.9: Valeurs de l'amélioration pour les différents aspects, moyennées sur tous les participants et experts.

Une différence significative est observée pour l'aspect évaluant l'utilisation de la pièce par les participants (*Stage usage*; $F(2, 132) = 3.627$, $p = 0.029$). Dans la condition utilisant l'audience virtuelle interactive ($\mu = 0.40$), la performance sur cet aspect s'améliore significativement plus par rapport à la condition de retour direct

Aspect	CTRL	RD	AVI
Regard	0.40 (1.37)	0.02 (1.32)	0.27 (1.27)
Posture	0.29 (1.12)	0.00 (1.13)	0.19 (1.12)
Débit de parole	0.16 (1.33)	0.17 (1.25)	0.40 (1.30)
Gestes	0.42 (1.39)	0.26 (1.15)	0.33 (1.24)
Intonation	0.29 (1.38)	-0.02 (1.09)	0.50 (1.35)
Confiance en soi	0.33 (1.49)	0.05 (1.45)	0.44 (1.58)
Utilisation de la pièce	0.42 (1.25)	-0.12 (0.99)	0.40 (0.89)
Hésitations	0.47 (1.01)	-0.07 (0.84)	0.35 (0.76)
Structure de la présentation	0.22 (1.35)	0.17 (1.38)	0.42 (1.15)
Performance globale	0.49 (1.42)	0.05 (1.45)	0.60 (1.32)

TABLE 8.1: Valeurs moyennes et écart-types (entre parenthèses) pour tous les aspects évalués par des experts dans les trois conditions, c'est à dire la condition de contrôle utilisant une audience passive (CTRL), la condition de retour direct (RD) et la condition utilisant l'audience virtuelle interactive (AVI).

($\mu = -0.12$; $t(88) = 0.94$, $p = 0.011$, $g = 0.543$). Une différence significative est aussi observée entre les participants dans la condition de contrôle ($\mu = 0.42$) et de retour direct ($\mu = -0.12$; $t(85) = 1.13$, $p = 0.029$, $g = 0.473$).

Pour l'évaluation de la quantité d'hésitations, nous observons une différence significative entre les conditions ($F(2, 132) = 4.550$, $p = 0.012$). Selon les experts, les participants assignés à la condition de l'audience interactive ($\mu = 0.35$; $t(88) = 0.80$, $p = 0.013$, $g = 0.530$) et les participants assignés à la condition de contrôle ($\mu = 0.47$; $t(85) = 0.93$, $p = 0.009$, $g = 0.572$) s'améliorent plus que les participants assignés à la condition de retour direct ($\mu = -0.07$).

Nous comparons enfin les améliorations globales des participants évaluées par les experts, en faisant la moyenne des améliorations (ou régressions) sur les dix aspects. On observe une différence significative entre les trois conditions ($F(2, 447) = 5.814$, $p = 0.003$; voir Figure 8.10). De manière générale l'amélioration entre les participants dans la condition utilisant l'audience interactive ($\mu = 0.39$, $\sigma = 0.83$; $t(298) = 0.86$, $p = 0.001$, $g = 0.395$) ou la condition de contrôle ($\mu = 0.35$, $\sigma = 1.05$; $t(288) = 0.98$, $p = 0.010$, $g = 0.305$) est significativement plus forte que dans la condition de retour direct ($\mu = 0.05$, $\sigma = 0.89$). Aucune différence significative n'est observée entre la condition de contrôle et la condition utilisant l'audience interactive.

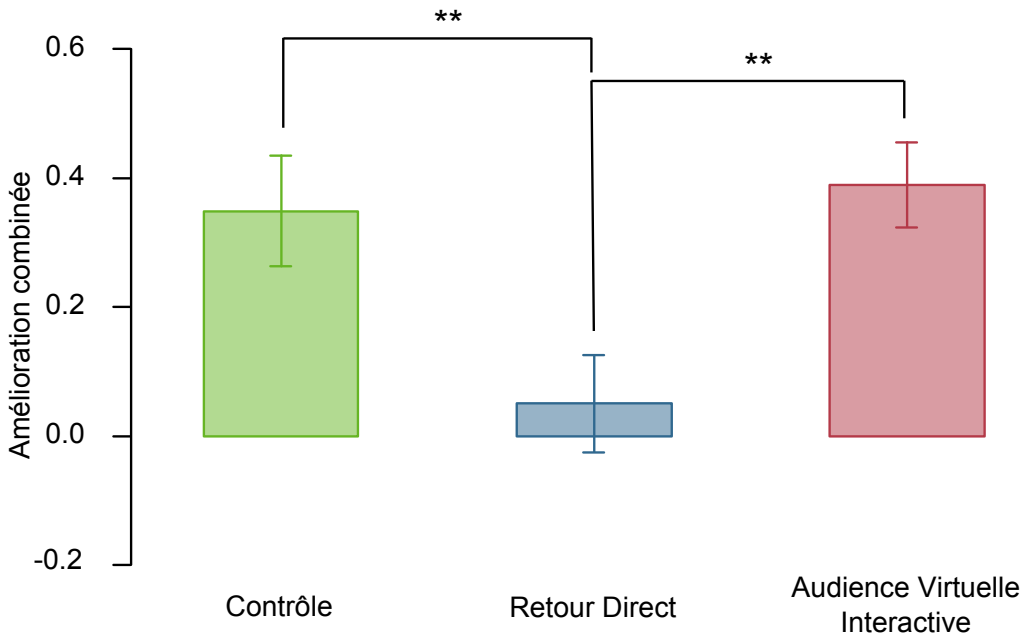


FIGURE 8.10: Visualisation de l'amélioration globale (moyenne des améliorations sur les dix aspects) évaluée par des experts, représentée avec la moyenne et écart-type. La condition de contrôle et la condition d'audience interactive sont significativement meilleures que la condition de retour direct ($** \rightarrow p < 0.01$)

8.4.3 Q3 - Évaluation objective

Ici, nous rapportons les différences observées dans les mesures objectives d'amélioration (voir Section 8.3.3.3), la quantité de regards orienté vers l'audience et la quantité d'hésitations, selon la condition d'entrainement. Nous réalisons une analyse de la variance à un facteur (*one-way ANOVA*) et des test T de Student bilatéraux entre les différentes conditions afin d'identifier si une des conditions entraine une amélioration plus grande d'un de ces deux aspects du comportement.

Nous n'observons pas de différence significative entre les trois conditions concernant la quantité de regards orientés vers l'audience ($F(2, 42) = 0.923$, $p = 0.405$). Cependant, dans les trois conditions, les participants ont amélioré leur comportement de regard (contrôle : $\mu = 0.21$, $\sigma = 0.29$; retour direct : $\mu = 0.10$, $\sigma = 0.19$; audience virtuelle interactive : $\mu = 0.12$, $\sigma = 0.18$).

La quantité d'hésitations a aussi été réduite pour toutes les conditions entre la présentation *pre-test* et la présentation *post-test* (contrôle $\mu = -0.39$, $\sigma = 0.35$;

retour direct $\mu = -0.36$, $\sigma = 0.35$; audience virtuelle interactive $\mu = -0.37$, $\sigma = 0.39$). Nous n'observons pas de différence significative entre les trois conditions ($F(2, 42) = 0.018$, $p = 0.982$).

8.5 Discussion

Dans cette section, nous discutons nos résultats par rapport aux trois perspectives introduites précédemment : les auto-évaluations des participants, les évaluations des experts, et l'évaluation objective.

8.5.1 Q1 - Auto-évaluation

Notre première perspective d'évaluation du système prend le point de vue des participants eux-mêmes. L'analyse des questionnaires a révélé plusieurs tendances dans les différentes conditions.

Tout d'abord, nous constatons que les participants ont trouvé que l'expérience comportait beaucoup de potentiel en termes de possibilités d'amélioration des compétences de prise de parole en public (SA_{Q29}). Les participants étaient de manière générale très intéressés par la possibilité de répéter l'expérience pour améliorer leurs compétences (SA_{Q30}). Même si les scores très hauts obtenus avec le système doivent être mis en perspective (*i.e.* il est probable qu'un biais positif soit dû à la nouveauté de l'expérience et à l'aspect imposant du système utilisant des personnages virtuels à taille réelle), ces résultats sont encourageants.

Deuxièmement, les participants ont trouvé l'interaction avec le système plutôt facile (SA_{Q17} et SA_{Q19}). En particulier, l'interaction a été jugée la plus facile par les participants interagissant avec le système configuré dans la condition de contrôle, ce qui n'est pas surprenant car les participants n'étaient gênés par aucun stimulus et n'avaient pas à se concentrer sur les retours du système (qu'ils soient directs ou indirects).

Enfin, les participants ont jugé le système très stimulant (SA_{Q2} and SA_{Q10}). Ils ont jugé l'audience interactive comme la plus engageante (SA_{Q1}). Ce résultat est prometteur, étant donné que des travaux précédents ont montré qu'une corrélation existe entre le degré d'engagement de l'apprenant d'une part, et la réussite du processus

d'apprentissage et la performance lors de tests postérieurs à l'apprentissage d'autre part [Rowe et al., 2010, Hart et al., 2013].

8.5.2 Q2 - Évaluations par les experts

Notre deuxième perspective d'évaluation se concentre sur les évaluations par les experts des vidéos des présentations *pre-test* et *post-test* des participants. Les experts évaluent l'amélioration du comportement des participants selon dix aspects faisant référence au comportement non-verbal, au comportement verbal, à la structure des présentations, et à des aspects généraux de la performance des participants. Nous avons eu recours à trois experts de la prise de parole en public provenant de l'association Toastmasters. Ces experts ont évalué les performances des participants à partir d'une interface où les vidéos *pre-test* et *post-test* étaient disposées l'une à côté de l'autre pour permettre une comparaison directe. Cette approche nous a permis de compenser le niveau initial de compétence de prise de parole en public des participants ainsi que l'opinion des experts sur un participant en particulier.

En général, nous observons que la performance globale des participants s'est améliorée pour les trois conditions. L'amélioration globale est la plus forte pour la condition utilisant l'audience virtuelle interactive, mais cet effet n'est pas significatif. En regroupant les différents aspects du comportement en une unique mesure, on observe que la condition utilisant l'audience virtuelle interactive et la condition de contrôle améliorent significativement plus le comportement des participants que la condition de retour direct (voir Figure 8.10).

De plus, nous observons des différences significatives sur certains aspects particuliers pour les trois conditions : l'intonation, l'utilisation de la pièce et la quantité d'hésitations se sont tous plus améliorés dans la condition utilisant l'audience virtuelle interactive qu'avec la condition de retour direct, tandis que l'utilisation de la pièce et la quantité d'hésitations se sont aussi plus améliorés dans la condition de contrôle que dans la condition de retour direct.

En conclusion, le système est prometteur pour améliorer les compétences de prise de parole en public pour tous les aspects évalués dans les conditions de contrôle et d'audience virtuelle interactive (voir Figure 8.9 et Table 8.1). Il semble cependant que la condition de retour direct montre une amélioration plus faible que les deux

Aspect	Expert 1	Expert 2	Expert 3
Regard	0.58	0.76	0.68
Posture	0.63	0.72	0.68
Débit de parole	0.74	0.86	0.71
Gestes	0.70	0.78	0.71
Intonation	0.66	0.92	0.70
Confiance en soi	0.83	0.89	0.81
Utilisation de la pièce	0.63	0.74	0.69
Hésitations	0.50	0.64	0.77
Structure de la présentation	0.73	0.50	0.85

TABLE 8.2: Coefficients de corrélation linéaire ρ de Pearson entre l'aspect de performance globale et les autres aspects du comportement.

autres conditions, parfois même négative pour certains aspects (utilisation de la pièce, hésitations). Ces stimuli additionnels (jauges colorées) ont pu constituer des distractions au lieu d'être des indices bénéfiques pour les participants. Inversement, le retour non-verbal produit par l'audience semble avoir été bénéfique, même si l'effet n'est pas significatif par rapport à la condition de contrôle.

Enfin, nous observons différentes préférences des experts, c'est à dire que certains experts attribuent plus d'importance à certains aspects du comportement qu'à d'autres dans la performance globale d'un participant. Nous avons étudié ce phénomène en calculant, pour chaque expert indépendamment des autres, les corrélations entre les évaluations de chaque aspect avec les évaluations de la performance globale des participants (voir Table 8.2). Par exemple, le troisième expert attribue une plus grande importance à la structure de la présentation ($\rho = 0.85$) que le deuxième expert ($\rho = 0.50$). Globalement, la confiance en soi est fortement corrélée avec la performance globale des participants.

8.5.3 Q3 - Évaluation objective

Nous évaluons enfin par le biais de mesures objectives si les participants ont amélioré certains aspects de leur comportement entre les présentations *pre-test* et *post-test*. Nous étudions la quantité de regards dirigés vers l'audience, et la quantité d'hésitations. Nous utilisons des annotations manuelles de ces comportements pour notre

évaluation.

Nous observons que les participants s'améliorent sur ces deux comportements de manière systématique (Quantité de regards : $\mu_{Contrôle} = 0.21$, $\mu_{RetourDirect} = 0.10$, $\mu_{AudienceInteractive} = 0.12$. Hésitations : $\mu_{Contrôle} = -0.39$, $\mu_{RetourDirect} = -0.36$, $\mu_{AudienceInteractive} = -0.37$). Cependant, nous n'observons pas de différence selon la condition dans laquelle se trouvaient les participants. Ceci suggère que porter l'attention des participants sur certains comportements peut leur permettre de s'améliorer indépendamment du type de retour produit par le système.

En dernier lieu, nous avons voulu étudier si des différences d'apprentissage existent entre des participants qui présentent différents niveaux préalables de compétences à la prise de parole en public. Nous observons ainsi que des personnes présentant initialement une bonne maitrise de la prise de parole en public ne tirent pas autant profit de l'entrainement avec le système. Cet effet de plafonnement apparait en particulier pour l'aspect de quantité de regards dirigés vers l'audience. Nous divisons la population en trois parties de taille équitables : faibles, moyens, et bons présentateurs, indépendamment de la condition d'apprentissage. Nous regroupons les participants en fonction de leur performance lors de la présentation *pre-test* : nous définissons les présentateurs faibles comme ceux qui dirigent leur regard le moins vers l'audience lors de la présentation *pre-test*. Globalement, nous observons une différence significative entre les trois groupes ($F(2, 42) = 36.762$, $p < 0.001$), et des tests T de Student réalisés pour comparaison les groupes deux à deux sont tous significatifs ($p < 0.01$) : les résultats sont rapportés en Figure 8.11.

Conclusion

Un système d'audience virtuelle pourrait présenter des avantages par rapport aux approches traditionnelles d'entrainement à la prise de parole en public. En effet, un tel système ne poserait pas de problèmes de disponibilité, permettrait un entrainement standardisé, et des personnes anxieuses lors de situations de prise de parole en public pourraient s'entrainer sans peur d'être jugées. Des audiences virtuelles ont déjà été utilisées pour réduire l'anxiété liée à la prise de parole en public. Jusque-là ces systèmes ne se sont pas directement intéressés à l'amélioration de la compétence des utilisateurs.

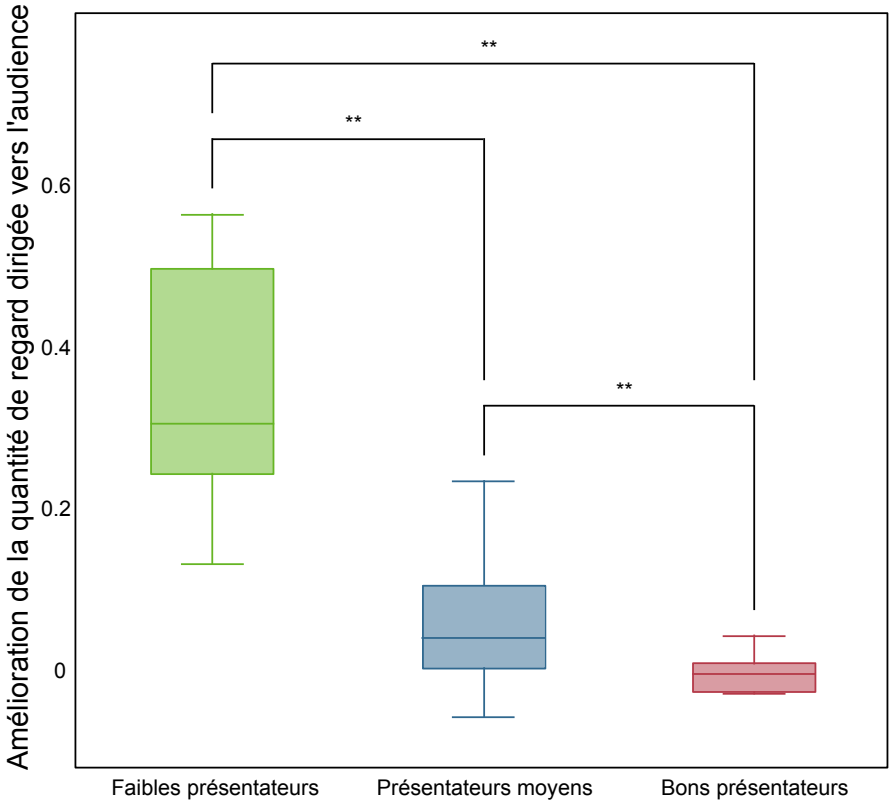


FIGURE 8.11: Plafonnement de l'amélioration du comportement de regard après l'apprentissage. Les trois groupes correspondent aux présentateurs faibles, moyens, et bons dans leur comportement de regard dans la présentation *pre-test* ($** \rightarrow p < 0.01$).

Dans ce chapitre, nous avons proposé une architecture pour un système d'entrainement à la prise de parole en public. Dans cette architecture, un ou plusieurs logiciels de perception multimodale (ou un *magicien d'Oz*) détectent les signaux de l'utilisateur, un module central agrège ces signaux et en dérive des indicateurs caractérisant la performance de l'utilisateur, et un retour à l'utilisateur sur sa performance peut être fourni par le biais d'éléments de visualisation directe ou par le comportement des membres de l'audience virtuelle.

Pour évaluer notre système, nous avons réalisé une étude visant à l'amélioration du comportement de regard et de la quantité d'hésitations. Dans cette étude, un *magicien d'Oz* était chargé de détecter les hésitations de l'utilisateur et sa direction de regard. Nous avons évalué trois différentes stratégies de retour à l'utilisateur, correspondant à (1) une audience virtuelle interactive, (2) des éléments de retour direct

et (3) une condition de contrôle utilisant une audience virtuelle passive. L'étude suivait un protocole *pre-test/post-test*, c'est à dire que les participants étaient enregistrés réalisant une présentation avant et après deux sessions d'entrainement avec le système. Nous avons évalué les résultats de l'étude selon trois perspectives : (Q1) l'auto-évaluation des présentateurs, (Q2) le jugement d'experts en prise de parole en public et (Q3) des mesures objectives de performance.

Nos résultats principaux sont les suivants : (Q1) les participants ont de manière générale apprécié utiliser le système et s'entrainer avec l'audience virtuelle, et ils ont trouvé que l'audience interactive était la plus stimulante et la plus engageante. (Q2) Les experts ont identifié une amélioration du comportement des présentateurs plus forte pour les conditions de contrôle et d'audience virtuelle interactive. Nous n'observons pas de différence significative entre ces deux conditions. (Q3) Une évaluation objective de deux comportements importants lors de prises de parole en public, la quantité de regards dirigés vers l'audience et la quantité d'hésitations, montrent une amélioration générale indépendamment de la condition, mais plus forte pour des personnes moins expérimentées.

Nous concluons de ces résultats que le système que nous avons proposé possède un fort potentiel pour l'entrainement des compétences de prise de parole en public, ce qui étend les résultats d'études passées ayant démontrées le potentiel de tels systèmes dans la réduction de l'anxiété liée à la prise de parole.

Synthèse du chapitre

1. L'approche traditionnelle pour l'amélioration de la compétence de prise de parole en public, qui repose sur un recours à des spécialistes ou à une audience de connaissances qui font pratiquer les apprenants et leur fournissent des retours, pose des problèmes de disponibilité, de standardisation et de réticence dans le cas de personnes sujettes à l'anxiété dans les situations de prise de parole en public.
2. Nous avons présenté une architecture d'audience virtuelle interactive qui permet de produire un retour en temps réel à l'utilisateur sur sa performance. Ce retour est réalisé par des éléments graphiques génériques ou par le comportement des personnages constituant l'auditoire.
3. Nous avons étudié trois configurations du système : une audience passive, une audience enrichie d'éléments graphiques donnant un retour direct, et une audience interactive fournissant un retour indirect par son comportement. Notre étude a réuni 45 participants. De manière générale, l'audience interactive a été trouvée la plus engageante et la plus motivante. Des experts ont jugé que les participants utilisant l'audience passive et interactive s'améliorent plus que ceux utilisant l'audience enrichie d'éléments de retour direct. Nous n'observons pas de différence significative entre l'audience passive et interactive.

9 Conclusion

Notre thèse s'inscrit dans le domaine des Agents Conversationnels Animés pour l'entrainement social. Nous avons contribué à deux enjeux de ce domaine. Notre contribution principale concerne la planification du comportement d'Agents Conversationnels Animés pour l'expression d'attitudes. Parallèlement, nous avons contribué à la problématique de la conception et de l'évaluation de systèmes d'entrainement de la compétence de prise de parole en public. Dans ce chapitre, nous résumons d'abord notre approche en rappelant les contributions de notre thèse. Nous présentons ensuite les limites que nous y avons identifiées. Enfin, nous terminons en dressant des perspectives de recherches futures.

9.1 Résumé de la thèse

9.1.1 Extraction et planification de séquences de signaux non-verbaux pour l'expression d'attitudes sociales

Les modèles de planification de comportement d'Agents Conversationnels Animés considèrent généralement les signaux non-verbaux de manière isolée. Or des travaux récents ont montré que le sens de certains signaux non-verbaux peut être altéré par d'autres signaux non-verbaux proches [Keltner, 1995, With & Kaiser, 2011, Jack et al., 2014]. De plus, les attitudes sociales ne s'expriment pas à un instant donné mais affectent le comportement sur de longues périodes [Scherer, 2005]. Certaines approches de planification de comportement ont considéré des séquences de signaux, dans le cadre de certains signaux communicatifs et de l'expression d'émotions (voir Section 3.3). Cependant, ces approches présentent certaines limites. Par exemple, celles-ci reposent souvent sur un travail de définition manuel des relations temporelles entre signaux d'une même séquence, ou ne considèrent qu'une seule modalité

à la fois. Nous nous sommes attachés à pallier ces limites en proposant un modèle de planification de séquences de signaux multimodaux reposant sur une méthode d'extraction automatique de séquences à partir d'un corpus multimodal. Nous avons appliqué notre modèle de planification de comportement à l'expression d'attitudes par un recruteur virtuel. En effet, un recruteur virtuel utilisé pour la simulation d'entretiens d'embauche doit pouvoir exprimer des attitudes sociales. Cependant, les modèles d'expression d'attitudes sociales existants comportent certaines limites (voir Section 3.2), notamment celle de ne pas considérer des séquences de signaux non-verbaux. Dans ce cadre, nous avons apporté les contributions suivantes.

- *Première contribution :* Annotation d'un corpus multimodal d'entretiens d'embauche au niveau des signaux non-verbaux et de l'expression d'attitudes.

Afin de pouvoir étudier le lien entre les séquences de signaux non-verbaux et l'expression d'attitudes, la première contribution de notre travail a été l'*annotation d'un corpus multimodal d'entretiens d'embauche*. Dans ce corpus, nous avons annoté les signaux non-verbaux que nous avons identifiés dans la littérature comme étant liés à l'expression d'attitudes sociales. Nous avons proposé l'utilisation d'outils d'annotation continue pour l'annotation des attitudes sociales, du fait de la représentation bi-dimensionnelle des attitudes [Argyle, 1988] et que celles-ci ne s'expriment pas à un instant donné mais tout au long d'une interaction [Scherer, 2005]. Nous avons tiré profit de cette nature continue afin de détecter automatiquement les instants de variation des attitudes exprimées par les recruteurs.

- *Deuxième contribution :* Méthodologie d'extraction de séquences de signaux non-verbaux caractéristiques de variations d'attitudes.

Nous avons ensuite proposé une *méthodologie d'extraction automatique de séquences de signaux non-verbaux*. Celle-ci consiste à regrouper les séquences de signaux non-verbaux précédant des variations d'attitudes en fonction du type de variation, puis à appliquer une méthode de fouille de motifs séquentiels sur les groupes de séquences de signaux pour chaque type de variation d'attitude. L'application de cette méthode à notre corpus multimodal nous a permis d'obtenir un ensemble de séquences de signaux non-verbaux observées fréquemment avant chaque type de variation d'attitude.

- *Troisième contribution :* Modèle de planification de séquences de signaux non-verbaux pour l'expression d'attitudes.

Ces séquences extraites par fouille de données ont ensuite été utilisées pour construire un *modèle de planification de séquences de signaux non-verbaux pour l'expression d'attitudes*. A partir d'une phrase enrichie d'intentions communicatives exprimée dans le format FML [Heylen et al., 2008], ce modèle calcule une séquence de signaux non-verbaux, exprimée dans le format BML [Vilhjálmsson et al., 2007]. Cette séquence exprime simultanément les intentions communicatives de la phrase et une attitude sociale. Le principe général de notre modèle est de générer de nombreuses *séquences candidates* de signaux non-verbaux. La première étape consiste à choisir des signaux non-verbaux exprimant les intentions communicatives de la phrase, puis à ajouter des signaux non-verbaux additionnels là où cela est possible. Ces deux étapes permettent d'obtenir un grand nombre de *séquences candidates*. Enfin, la dernière étape consiste à sélectionner la séquence la plus susceptible d'exprimer l'attitude voulue parmi ces *séquences candidates*.

- *Quatrième contribution :* Architecture, implémentation et évaluation d'un recruteur virtuel autonome.

Nous avons implémenté un *recruteur virtuel autonome*, dont l'architecture est centrée sur notre modèle de planification de séquences. Nous avons évalué l'expression d'attitudes par ce recruteur virtuel lors de simulations d'entretiens d'embauche. Cette étude a permis de confirmer que les attitudes sont les mieux reconnues lorsque les modalités verbales et non-verbales participent à cette expression simultanément. De plus, nous avons observé que des participants évaluant le recruteur par le biais de vidéos reconnaissent mieux les attitudes exprimées.

9.1.2 Une audience virtuelle interactive pour l'entrainement à la prise de parole en public

La deuxième problématique de notre travail a été d'étudier le potentiel qu'auraient des audiences virtuelles pour l'amélioration de la compétence de prise de parole en public.

- *Cinquième contribution :* Architecture et implémentation d'une audience virtuelle interactive.

Notre première contribution dans ce domaine est d'avoir proposé une *architecture d'audience virtuelle pour l'entrainement de compétence de prise de parole en public*

permettant de fournir un retour en temps réel à l'utilisateur sur sa performance par le biais de signaux non-verbaux produits par les ACAs constituant l'audience. Ce type de système est, à notre connaissance, le premier à donner un retour en temps réel à l'utilisateur sur la qualité de son comportement lors de séances d'entrainement de la compétence de prise de parole en public.

- *Sixième contribution :* Comparaison de stratégies de retour en temps réel à l'utilisateur.

Nous avons évalué trois différentes stratégies de retour sur la performance de l'utilisateur : une audience passive, une audience augmentée par des éléments graphiques donnant une information directe sur la performance de l'utilisateur, et une audience interactive fournissant un retour à l'utilisateur par des signaux non-verbaux. A cette fin, nous avons utilisé trois perspectives d'évaluation : l'avis des utilisateurs eux-mêmes, celui d'experts évaluant l'amélioration des utilisateurs de système, et des mesures objectives d'amélioration. L'audience augmentée par des éléments graphiques a été la moins bonne selon les trois perspectives. Les audiences passives et interactives n'ont pas pu être départagées par les avis d'experts et les mesures objectives. Cependant, l'audience interactive a recueilli de meilleures impressions de la part des participants.

9.2 Limites

Nos contributions au domaine des Agents Conversationnels Animés pour l'entrainement social ne sont cependant pas exemptes de limites. Nous présentons ici celles que nous avons identifiées.

9.2.1 Corpus multimodal pour l'étude de l'expression d'attitudes

Lorsque nous avons analysé les corpus multimodaux existants potentiellement réutilisables pour notre travail, nous n'avons pas identifié de corpus disponible répondant parfaitement à nos besoins. Nous avons donc utilisé des vidéos d'entretiens d'embauche collectées dans le cadre du projet Tardis. Les conditions d'enregistrement de ces vidéos n'étaient pas idéales (*i.e.* les scènes étaient filmées de côté). Ceci a impliqué un certain nombre de contraintes : une annotation manuelle a dû être réalisée,

moins précise qu'elle aurait pu l'être avec de meilleures vidéos. L'annotation de l'attitude aurait aussi gagné à être réalisée par plus de participants. Ces contraintes ont eu un impact sur la qualité du corpus. L'accord inter-annotateurs sur les attitudes, notamment, s'est révélé moyen. Notre travail aurait naturellement pu bénéficier d'un corpus plus conséquent.

Ultérieurement à la réalisation de nos propres travaux, le corpus HuComTech [Hunyadi et al., 2012] est devenu disponible, comprenant un grand nombre d'entretiens, et une annotation de certaines modalités non-verbales et verbales. Ce corpus pourra être utilisé pour valider notre méthodologie d'extraction automatique de séquences de signaux non-verbaux dans un autre contexte.

9.2.2 Représentation des séquences

A notre connaissance, notre travail constitue la première tentative d'utiliser directement des séquences de signaux non-verbaux extraites par fouille de données afin de planifier le comportement d'un Agent Conversationnel Animé. Nous avons choisi une représentation des séquences modélisant uniquement la relation d'ordre entre plusieurs signaux, afin de simplifier la planification de séquences. Cela implique que nous ne modélisons pas les relations temporelles précises entre signaux d'une séquence. Or, il est possible que le sens d'une séquence soit modifié par ces relations temporelles.

Des techniques de fouilles de données plus complexes ont été proposées, permettant en particulier d'extraire des motifs où sont représentés des relations temporelles entre des évènements [Antunes & Oliveira, 2001, Guillame-Bert & Crowley, 2012]. L'utilisation d'une technique de fouille de motifs séquentiels plus complexe aurait cependant accentué les faiblesses du corpus multimodal rapportées plus haut. De plus, des représentations séquentielles plus complexes auraient complexifié la tâche de planification d'une séquence de comportement synchronisée avec le discours.

9.2.3 Modèle de planification de comportement pour l'expression d'attitudes

La limite principale de notre modèle de planification de séquences de signaux non-verbaux est que celui-ci échoue à exprimer l'attitude de soumission. En effet, lors de l'évaluation du modèle (voir Section 6.3), nous avons observé que l'attitude de soumission n'était pas reconnue pas les participants de l'étude. Nous pensons que plusieurs aspects du contexte de l'étude ont pu contribuer à ce résultat négatif. Tout d'abord, le recruteur est naturellement la personne qui domine un entretien d'embauche : c'est lui qui choisit la structure de l'entretien, et qui a le pouvoir d'accorder ou de refuser le poste au candidat. De plus, le simple fait de prendre la parole revient à prendre le contrôle sur l'interaction, c'est à dire une expression de dominance.

Notre modèle ne parvient pas non plus à exprimer des attitudes de diverses intensités. Nous supposons que l'absence de modulation d'expressivité dans le comportement lors de notre étude préliminaire a pu influencer ce résultat. Une évaluation pourrait être réalisée afin d'étudier si l'ajout du modulateur d'expressivité décrit dans la section 7.1.2.4 permet de pallier cette limite.

Une autre limite de notre modèle est que celui-ci ne prend en compte qu'une seule dimension à la fois lors de la planification de comportement, bien que la représentation des attitudes d'Argyle que nous utilisons soit bi-dimensionnelle. Or, il est possible, par exemple, qu'une séquence de signaux non-verbaux exprime simultanément de l'amicalité et de la soumission, ou simultanément de l'hostilité et de la dominance.

Enfin, la première étape de notre modèle de planification de séquences consiste à construire des séquences comprenant un signal par intention communicative contenue dans le fichier FML reçu en entrée (voir Section 6.2.1). Nous avons choisi de ne considérer qu'un seul signal par intention communicative car aucune séquence fréquente extraite par notre méthodologie de fouille de données ne comprenait de combinaison de signaux. Nous ne pourrions donc pas évaluer des combinaisons de signaux lors de l'étape de classification (Section 6.2.3). Or, il est tout à fait possible que plusieurs signaux multimodaux soient utilisés simultanément pour réaliser une intention communicative, et des modèles de planification existants de comportement

prennent ce phénomène en compte [Mancini & Pelachaud, 2007].

9.2.4 Recruteur virtuel autonome

Notre architecture de recruteur virtuel autonome comporte plusieurs composants : un planificateur du comportement d'écoute, un modèle de dialogue, notre planificateur de séquences de comportements non-verbaux (enrichi d'un modulateur d'expressivité) et un réalisateur de comportements. Nous n'avons pas considéré l'influence sur l'expression d'attitudes du planificateur du comportement d'écoute. Or, la réalisation du tour de parole et les signaux produits en écoutant un interlocuteur peuvent participer à l'expression d'attitudes. Si ce composant a été réalisé de manière à être le plus neutre possible, la réalisation du tour de parole et du comportement d'écoute ont pu avoir un impact sur la perception de l'attitude du recruteur virtuel.

9.2.5 Audience virtuelle interactive

Notre implémentation d'audience virtuelle comporte un certain nombre de limites techniques : les seuls modèles de personnages virtuels dont nous disposions étaient d'apparence similaire et de sexe masculin. Certains utilisateurs du système nous ont indiqué qu'une audience plus diversifiée serait préférable. De plus, lors de notre étude, les logiciels de détection de comportement multimodaux de l'utilisateur n'était pas encore intégrés à notre système. C'est donc un *magicien d'Oz* qui a indiqué au système la performance de l'utilisateur, et la perception du comportement des utilisateurs a ainsi pu être biaisée.

Une autre limite de ce travail est que nous n'avons pas validé le choix des comportements positifs et négatifs de l'audience virtuelle interactive. Il se peut que les changements de posture et mouvements de tête ne soient pas les signaux plus adaptés pour fournir un retour à l'utilisateur sur sa performance.

9.3 PERSPECTIVES

Nous envisageons plusieurs perspectives de travail principales. Certaines de ces perspectives permettraient à court-terme de renforcer nos contributions en palliant cer-

taines des limites que nous avons présentées ci-dessus. Les autres constituent de nouvelles questions de recherche qui nous sont apparues au fur et à mesure de notre travail.

9.3.1 Perspectives à court-terme

9.3.1.1 Planification de séquences contenant des combinaisons de signaux multimodaux

Une des limites de notre modèle de planification de séquences de signaux non-verbaux pour l'expression d'attitudes est que les intentions communicatives à exprimer ne sont réalisées que par un seul signal non-verbal à la fois. Nous proposons de modifier l'étape de classification des *séquences candidates* de notre méthode afin de pouvoir évaluer des séquences contenant des combinaisons de signaux non-verbaux, même si les séquences fréquentes qu'elle utilise dans le processus de classification ne contiennent pas de telles combinaisons.

Par exemple, lors de l'étape d'extraction des sous-séquences de la *séquence candidate* $< HeadDown, EyebrowUp >\rightarrow GestComm \rightarrow HeadAt$, nous pourrions considérer les signaux simultanés séparément et extraire les sous-séquences de la séquence $HeadDown \rightarrow GestComm \rightarrow HeadAt$ et de la séquence $EyebrowUp \rightarrow GestComm \rightarrow HeadAt$, en s'assurant qu'aucune sous-séquence n'est extraite en doublon.

9.3.1.2 Expression d'attitudes bi-dimensionnelles

Une autre limite de notre modèle d'expression d'attitudes par le biais de séquences de signaux non-verbaux est qu'il ne considère qu'une seule dimension du modèle d'Argyle à la fois. Or, lorsqu'un ACA utilise une séquence de signaux non-verbaux pour exprimer, par exemple, de l'amicalité, il est tout à fait possible que cette séquence ait aussi un effet sur la dominance exprimée.

Nous envisageons d'adapter notre modèle de planification de séquences de signaux non-verbaux afin de considérer simultanément les deux dimensions de l'attitude.

Cette adaptation pourrait avoir lieu au niveau de la dernière étape de notre algorithme (*i.e.* sélection de la séquence finale). Dans le modèle que nous avons proposé, la séquence choisie est celle qui contient le plus de sous-séquences exprimant la variation voulue de la dimension d'attitude considérée (vote à la majorité). Nous pourrions modifier cette étape afin que la séquence choisie soit celle contenant le plus de sous-séquences exprimant la variation voulue pour les deux dimensions d'attitudes.

9.3.1.3 Amélioration du système d'audience virtuelle interactive et étude des types de signaux produits par l'audience

Nous envisageons aussi plusieurs perspectives d'amélioration de notre système d'audience virtuelle interactive pour l'amélioration de la compétence de prise de parole en public. Tout d'abord, nous allons modifier notre système en tenant compte des limites que nous avons présentées plus haut : l'audience virtuelle sera diversifiée, et nous allons achever l'intégration des composants logiciels permettant de détecter les comportements de l'utilisateur en temps réel et d'en déduire automatiquement des mesures de performance.

En outre, nous proposons d'étudier l'influence des types de signaux produits par l'audience virtuelle sur l'amélioration de la compétence des utilisateurs. En effet, nous avons considéré des comportement positifs et négatifs naturellement observés dans des audiences réelles mais relativement subtils de la part de membres d'une large audience (*i.e.* changements de posture, hochements de tête). Des mouvements plus exagérés (*ex.* endormissements, sorties de pièce, applaudissements) pourraient peut-être se révéler plus efficaces.

9.3.2 Perspectives à long-terme

9.3.2.1 Motifs temporels complexes pour la modélisation du comportement d'écoute et du tour de parole

Le modèle que nous avons proposé permet à un ACA d'exprimer des attitudes lorsqu'il prend la parole. Cependant, l'influence de l'attitude sur le comportement

d'écoute et de tour de parole n'est pas modélisée. Une perspective pour améliorer l'expression de l'attitude par le recruteur virtuel serait de considérer des modèles de comportement d'écoute et de tour de parole qui prennent en compte les attitudes sociales. Pour ces deux aspects du comportement, la synchronisation temporelle avec l'interlocuteur est un aspect très important [Allwood et al., 2007b].

La représentation séquentielle que nous avons utilisée pour la planification du comportement multimodal lors de la prise de parole est relativement simple et ne permettrait pas de représenter cette synchronisation temporelle de manière adéquate. Cependant, des modèles plus complexes existent et pourraient être adaptés à la modélisation du comportement d'écoute et de prise de parole. Par exemple, le modèle proposé par Guillame-Bert permet d'extraire des motifs représentant les relations temporelles complexes entre des évènements symboliques [Guillame-Bert & Crowley, 2012]. Celui-ci peut notamment modéliser l'incertitude et les distributions temporelles des évènements d'un motif (*ex.* si l'évènement A arrive à l'instant t, alors l'évènement B a 95% de chances d'arriver entre t et $t+5$; dans ce cas, la distribution temporelle de B entre t et $t+5$ suit une loi gaussienne). L'extraction de motifs plus complexes nécessitera cependant d'utiliser des corpus multimodaux plus précis et plus conséquents.

9.3.2.2 Étude longitudinale de l'influence d'une audience virtuelle sur la compétence de prise de parole en public

Enfin, nous nous proposons de vérifier que l'amélioration de la compétence de prise de parole en public après l'entrainement avec une audience virtuelle constatée lors de notre étude se transfère dans des conditions réelles. Afin d'étudier cela, nous pourrions utiliser des audiences réelles pour évaluer les participants avant et après des séquences d'entrainement.

De plus, nous comptons espacer les présentations d'évaluation et d'entrainement dans le temps. En effet, dans notre étude, les participants réalisaient l'évaluation *post-test* immédiatement après l'entrainement, qui était lui-même réalisé immédiatement après l'évaluation *pre-test*. Nous pouvons nous demander si l'amélioration que nous avons constatée est durable (*i.e.* six mois après l'utilisation du système, la compétence de prise de parole en public a-t-elle régressé à son niveau original). Par

ailleurs, l'étude que nous avons réalisée était assez intensive (une heure par participant), ce qui a pu fatiguer les participants et influencer leur performance finale. Nous envisageons donc de réaliser une étude longitudinale afin d'évaluer l'effet d'un entrainement à la prise de parole en public avec une audience virtuelle étalé sur de plus longues périodes.

A

Documents de l'évaluation du recruteur virtuel

Cette annexe est consacrée à l'étude ayant pour but d'évaluer le recruteur virtuel, en interaction et par des vidéos, décrite au chapitre 7.

Le premier document (page 190) est le formulaire de consentement que devaient signer les participants afin de pouvoir participer à l'étude.

Le deuxième document (page 191) est une copie d'écran des instructions affichées sur l'ordinateur utilisé par les participants réalisant l'évaluation *en interaction*. Le troisième document (page 192) est quant à lui une copie d'écran des instructions pour les participants réalisant l'évaluation par le biais de *vidéos*.

La section A.3 présente l'ensemble des actes de dialogues et phrases correspondantes du modèle de dialogue dans l'évaluation du recruteur virtuel.

Les tableaux présentés aux pages 199-202 rassemblent les moyennes des réponses données par les participants à l'étude. Le premier tableau regroupe les réponses des participants évaluant le recruteur virtuel en interaction. Le deuxième tableau regroupe les réponses des participants évaluant le recruteur virtuel par des vidéos. Le troisième tableau regroupe ces deux catégories ensemble.

A.1 DÉCLARATION DE CONSENTEMENT POUR L'ÉVALUATION DU RECRUTEUR VIRTUEL EN INTERACTION

Evaluation TARDIS

DECLARATION DE CONSENTEMENT

Novembre 2014

Chercheur responsable:
Mathieu Chollet, CNRS-LTCI, Telecom ParisTech

Dans cette étude, intitulée "Evaluation TARDIS", vous allez participer à un entretien d'embauche virtuel. Votre rôle consiste à jouer le rôle d'un candidat à une offre d'emploi. Le recruteur va vous poser de questions, et vous être libre d'y répondre comme bon vous semble.

Aucune donnée audio ou vidéo n'est enregistrée. Les seules données conservées de votre participation sont les informations entrées lors de questionnaires web pour évaluer le personnage virtuel. Ces informations sont anonymes et ne permettent pas de vous identifier.

Je comprends que ma participation est volontaire, que je suis libre de retirer mon consentement à participer et arrêter l'expérience *à tout moment* sans conséquence pour moi. La procédure expérimentale m'a été expliquée et le chercheur responsable de l'expérience, Mathieu Chollet, m'a proposé de répondre à toute question à propos de celle-ci.

J'ai lu et compris les informations ci-dessus et accepte de participer dans cette étude, et permet aux chercheurs de conserver les données collectées lors de ma participation.

_____
NOM COMPLET DU PARTICIPANT

_____
SIGNATURE DATE

A.2 CAPTURES D'ÉCRANS DES QUESTIONNAIRES SUR NAVIGATEUR POUR L'ÉVALUATION DU RECRUTEUR VIRTUEL

Evaluation du recruteur virtuel

Veuillez indiquer les informations suivantes :

Genre: M ○ F ○
Age:
Nationalité:

Cette étude est menée à des fins de recherches académiques. L'étude est anonyme : nous ne vous demanderons pas de fournir d'informations permettant de vous identifier, comme votre nom. L'expérience prend une vingtaine de minutes à réaliser.

Dans cette étude, vous jouez le rôle d'un candidat à une offre d'emploi. Vous allez rencontrer un recruteur avec lesquel vous aurez une courte conversation. Le but de l'expérience est d'évaluer les modèles de comportements pour des personnages virtuels. **Important :** Si le recruteur vous entend, ils ne comprend pas ce que vous dites. Il n'y a pas de bonne ou de mauvaise réponse. Lors de vos réponses aux questions du recruteur, vous pouvez tout à fait vous inventer un personnage, ou vous inspirer de votre vécu. L'important est de vous sentir à l'aise.

L'entretien avec le recruteur est découpé en quatre phases, où le recruteur vous posera des question sur des sujets distincts. Le premier personnage est la secrétaire de l'entreprise : celle-ci va vous expliquer comment les entretiens vont se dérouler. Ce personnage est aussi ici pour vous donner l'occasion de vous habituer à interagir avec un personnage virtuel. Après cela, vous verrez le recruteur. D'abord, celui-ci va vous demander de vous présenter de manière générale. Lors de la deuxième phase, il vous posera des questions sur votre formation et vos études. Dans la troisème phase, vous parlerez de votre expérience professionnelle. Enfin, vous discuterez dans la quatrime phase de vos compétences.

Entre chaque phase, nous vous demanderons de remplir un questionnaire afin d'obtenir l'impression que le recruteur vous a laissé durant cette phase.

Lorsque vous êtes prêt a démarrer, merci de prévenir le chercheur en charge de l'expérience. C'est aussi un moment idéal pour poser des questions si vous en avez.

Envoyer

Evaluation du recruteur virtuel

Veuillez indiquer les informations suivantes :

Genre: M ○ F ○
Age:
Nationalité:

Cette étude est menée à des fins de recherches académiques. L'étude est anonyme : nous ne vous demanderons pas de fournir d'informations permettant de vous identifier, comme votre nom. L'étude prend une quinzaine de minutes au total.

Important : vous devez pouvoir entendre le recruteur : si ce n'est pas déjà le cas, merci de vous assurer que le son fonctionne sur votre ordinateur.

Dans cette étude, vous allez voir un recruteur poser douze questions. Le but de l'expérience est d'évaluer les modèles de comportements de personnages virtuels.

Entre chaque phrase du recruteur, nous vous demandons de remplir un questionnaire afin d'obtenir l'impression que celui-ci vous a laissé en prononçant cette phrase, de par le choix de ses mots ou son comportement : expressions faciales, gestes, etc...

Lorsque vous êtes prêt a démarrer, merci de cliquer sur le bouton ci-dessous.

Envoyer

A.3 PHRASES ET ACTES DE DIALOGUE DE L'ÉVALUATION

Dans cette section, nous rapportons l'ensemble des actes de dialogue et phrases correspondantes prononcées par les agents de l'évaluation du recruteur virtuel. L'ordre de présentation de ces actes de dialogue suit celui du scénario de l'étude (*ex.* dans la deuxième phase, le modèle de dialogue commençait par choisir l'acte de dialogue *Welcome*, puis *AskAcademicInterest*, puis *AskImportanceStudies*, *etc.*)

Phase d'habitation

Habituation-Welcome :

Bonjour, bienvenue. Je suis Laure, et je suis un agent virtuel ! Ici, nous entrainons les gens à passer des entretiens d'embauches. Je vous en dirai plus tout à l'heure, mais d'abord, comme je ne sais pas si vous avez déjà parlé avec un personnage virtuel, je vais commencer par vous expliquer certaines choses importantes. D'accord ?

Habituation-SpeechRecognition :

Tout d'abord, vous devez savoir que même si nous pouvons vous entendre, nous, les agents virtuels, ne comprenons pas ce que vous dites ! Du coup, vous pouvez vous entrainer avec nous. Il n'y a pas à s'inquiéter des conséquences ! Cela peut être amusant, vous ne trouvez pas ?

Habituation-ButExperienceAnonymat :

Les chercheurs de ce laboratoire font cette expérience afin de nous améliorer. Du coup, après avoir parlé avec mes collègues, nous vous ferons remplir un questionnaire pour savoir ce que vous avez pensé d'eux. Ces questionnaires sont anonymes et ce sont les seules informations que nous allons conserver. Nous n'enregistrons rien pendant cette expérience. OK ?

Habituation-Deroulement :

Bien, je vais vous expliquer comment cet entretien virtuel va se passer. Vous allez maintenant rencontrer ma collègue recruteuse. Celle-ci va vous poser des questions sur des sujets différents. Elle va commencer par vous demander de vous présenter et va vous poser des questions générales. Ensuite, vous allez parler de votre formation

et des études que vous avez faites. Dans la troisième partie, vous allez parler de votre expérience professionnelle. Enfin, vous terminerez par discuter de vos compétences interpersonnelles. Alors, est-ce que vous vous sentez prêt ?

Habituation-Goodbye :

C'est tout pour ma part. Vous allez maintenant rencontrer ma collègue. Mais avant cela, n'hésitez pas à poser des questions au chercheur en charge de l'expérience. Merci encore, amusez vous bien !

Première phase : questions générales

1stPhase-Welcome :

Bonjour ! Ensemble, nous allons évoquer votre profil de manière générale.

1stPhase-AskSelfGeneralDescription :

Neutre - Pour commencer, dites moi comment vous vous décriveriez.

Dominante - Pour commencer, vous allez vous décrire de manière générale. Allez-y.

Amicalité - Pour commencer, parlez moi de vous.

Inamicalité - Pour commencer, dites-moi comment vos collègues vous décriraient.

1stPhase-AskGenericCareerGoals :

Neutre - Comment voyez-vous votre trajectoire professionnelle ?

Dominante - Vous devez avoir des objectifs professionnels. Parlez en moi.

Amicalité - Comment vous voyez vous atteindre vos objectifs professionnels ?

Inamicalité - Comment envisagez vous votre carrière professionnelle dans le moyen terme, et dans le long terme ?

1stPhase-AskMotivations :

Neutre - Qu'est-ce qui vous motive au travail ?

Dominante - Parlez moi de vos motivations dans votre travail.

Amicalité - Quelles sont les choses qui vous passionnent au travail ?

Inamicalité - Décrivez vos tâches préférées dans votre travail.

1stPhase-Goodbye :

C'est tout pour l'instant. Veuillez maintenant remplir le questionnaire sur l'autre ordinateur. Ensuite, nous continuerons cet entretien.

Deuxième phase : formation et études

2ndPhase-Welcome :

Vous revoilà ! Maintenant, nous allons parler de votre formation et de vos études.

2ndPhase-AskAcademicInterest :

Neutre - Pourquoi avez-vous fait des études dans votre spécialité ?

Dominante - Maintenant, vous allez me dire pourquoi vous avez fait des études dans votre spécialité.

Amicalité - Qu'est-ce qui vous a plu le plus dans votre spécialité à l'université ?

Inamicalité - Pour quelle raisons avez vous choisi votre spécialité à l'université ?

2ndPhase-AskImportanceStudies :

Neutre - Décrivez l'importance que vous portez à vos études.

Dominante - Maintenant, vous allez me décrire l'importance que vous portez à vos études.

Amicalité - Est-ce que vos études ont été importantes pour vous ? Et pourquoi ?

Inamicalité - Quel impact ont eu vos études sur votre vie ?

2ndPhase-AskParallelStudies :

Neutre - Que faisiez vous pendant votre temps libre lorsque vous étiez encore étudiant ?

Dominante - Maintenant, décrivez moi vos activités pendant votre temps libre ?

Amicalité - Durant vos études, que faisiez vous lorsque vous n'étiez pas en cours ?

Inamicalité - Lors de vos années d'études, quelles étaient vos activités pendant votre temps libre ?

2ndPhase-Goodbye :

C'est tout pour l'instant. Veuillez maintenant remplir le questionnaire sur l'autre ordinateur. Ensuite, nous continuerons cet entretien.

Troisième phase : formation et études

3rdPhase-Welcome :

Rebonjour ! A présent, nous allons parler de votre expérience professionnelle.

3rdPhase-AskPreviousJobPositionDetails :

Neutre - Quelles étaient vos fonctions dans votre précédent travail ?

Dominante - Présentez moi les responsabilités que vous assumiez dans votre précédent travail.

Amicalité - Que faisiez vous dans votre précédent travail ?

Inamicalité - Décrivez vos responsabilités principales dans votre précédent travail.

3rdPhase-AskPreviousJobOpinion :

Neutre - Dans votre travail, qu'est-ce que vous aimiez et qu'est-ce que vous n'aimiez pas ?

Dominante - Parlez moi de ce que vous avez aimé dans votre précédent travail.

Amicalité - Qu'est-ce qui était le plus et le moins rétribuant dans votre précédent travail ?

Inamicalité - Décrivez les mauvais et bons aspects de votre précédent travail.

3rdPhase-AskPreviousJobSuccess :

Neutre - Quel a été votre plus grand accomplissement dans votre précédent emploi ?

Dominante - A présent, vous allez me décrire une situation dans votre précédent emploi où vous avez été très performant.

Amicalité - Quel a été votre plus grand succès ? Dans votre précédent travail.

Inamicalité - Décrivez un moment où vous avez fait preuve d'une haute performance dans votre précédent emploi.

3rdPhase-Goodbye :

C'est tout pour l'instant. Veuillez maintenant remplir le questionnaire sur l'autre ordinateur. Ensuite, nous continuerons cet entretien.

Quatrième phase : formation et études

4thPhase-Welcome :

Rebonjour ! Pour cette dernière phase, nous allons parler de vos compétences interpersonnelles.

4thPhase-AskTeamWork :

Neutre - Pouvez-vous travailler en équipe ?

Dominante - Dites moi si vous préférez travailler seul ou en équipe, et pourquoi.

Amicalité - Êtes-vous capable de travailler en équipe ?

Inamicalité - Est-ce que vous préférez travailler seul ou avec d'autres personnes ?

4thPhase-AskResilience :

Neutre - Comment gérez vous le stress ?

Dominante - Parlez moi de votre manière de gérer le stress au travail.

Amicalité - Comment faites-vous pour gérer le stress ?

Inamicalité - Comment vous comportez vous sous le stress et la pression ?

4thPhase-AskFlexibility :

Neutre - Décrivez comment vous vous êtes adapté aux nouvelles tâches de votre précédent travail.

Dominante - Maintenant, vous allez décrivre une situation où vous avez dû vous adapter à une nouvelle tâche.

Amicalité - Parlez-moi d'un moment où vous vous êtes adapté à une nouvelle tâche

Inamicalité - Décrivez une situation où l'on vous a demandé de réaliser une tâche que vous n'aviez jamais réalisée précédemment.

4thPhase-Goodbye :

C'est tout pour ma part. Veuillez maintenant remplir le questionnaire sur l'autre ordinateur. Ensuite, cet entretien sera terminé. Merci encore !

A.4 RÉSULTATS DE L'ÉVALUATION DU RECRUTEUR VIRTUEL

Interaction													
	Amicalité				Inamicalité				Dominance				
	SPCH	MM	NTR	NVB	SPCH	MM	NTR	NVB	SPCH	MM	NTR	NVB
Q1	3,5	2,75	3,625	3,875	3,875	4,625	4,25	3,875	3	4,875	4,25	3,375
Q2	4	4,875	4,625	4,75	4,875	3,625	4,625	4,625	4,75	4,125	4,75	5,125
Q3	3,875	4,875	4,5	3,875	5,125	3,625	4,375	4,75	5,125	4	4,75	4,875
Q4	3,875	4,75	4,25	4,375	4,5	3	4,25	4	4,375	3,5	4,375	4,625
Q5	5	4,875	4,625	5,625	6,25	5,875	5,625	6,25	4,75	5,25	5,125	5,25
Q6	4,5	4,625	4,5	5,5	6	6,25	6,125	6,375	4,75	5,25	4,875	4,5
Q7	4	4,375	3,75	4,5	4,75	4,25	4,625	4,5	4,125	4	4,125	4,375
Q8	1,75	2	1,625	1,875	1,625	2,25	1,75	1,75	2	2,5	2,5	2,125
Q9	3,75	4,5	3,375	3,5	4,625	4,25	4,125	4,25	5,125	4,375	4,625	4,75
Q10	1,875	1,75	1,75	1,625	2	2,25	2,25	2,125	2,375	3,75	2,625	2,875

TABLE A.1: Moyennes des mesures pour les participants en interaction.

Vidéos												
	Amicalité				Inamicalité				Dominance			
	SPCH	MM	NTR	NVB	SPCH	MM	NTR	NVB	SPCH	MM	NTR	NVB
Q1	3,5	3,625	4,125	3,5	3,875	4,25	4,625	4,125	4,625	5,125	4,75	4,75
Q2	4,25	4,875	4,125	4	4,125	4	3,875	3,875	4,5	3,625	4,25	4
Q3	4,25	5	4,5	4,5	4,5	4,125	3,625	4,125	4,5	3,875	4,375	4,25
Q4	4	4,75	4	4,125	4	3,75	3,625	3,625	4,125	3,5	4,125	3,625
Q5	4,75	5,25	5	4,625	4,625	4,75	4,875	4,625	5,5	5,625	5,5	5,375
Q6	5	5,25	5,25	5,125	4,375	4,375	4,875	4,5	5,5	5,75	5,375	5,25
Q7	3,875	4,5	3,875	4,375	4,125	4,125	3,875	4	4,75	4	4,125	3,875
Q8	3,125	2,5	2,875	2,875	2,625	2,75	2,875	2,75	2,5	3,375	3,25	2,75
Q9	4,5	4,5	3,625	4,375	4,25	4,125	3,625	4,125	4,25	3,875	4	4,25
Q10	3	2,75	2,75	2,75	2,5	2,75	2,875	2,5	2,375	3,625	3,625	3

TABLE A.2: Moyennes des mesures pour les participants évaluant des vidéos.

Vidéos et en interaction												
	Amicalité				Inamicalité				Dominance			
	SPCH	MM	NTR	NVB	SPCH	MM	NTR	NVB	SPCH	MM	NTR	NVB
Q1	3,5	3,25	3,875	3,75	3,875	4,375	4,375	4	3,875	5	4,5	4
Q2	4,125	4,875	4,375	4,375	4,5	3,875	4,25	4,25	4,625	3,875	4,5	4,625
Q3	4,125	4,875	4,5	4,25	4,875	3,875	4	4,5	4,875	4	4,625	4,625
Q4	4	4,75	4,125	4,25	4,25	3,375	3,875	3,875	4,25	3,5	4,25	4,125
Q5	4,875	5	4,875	5,125	5,375	5,375	5,25	5,5	5,125	5,375	5,375	5,375
Q6	4,75	5	4,875	5,375	5,25	5,25	5,5	5,5	5,125	5,5	5,125	4,875
Q7	4	4,5	3,75	4,5	4,375	4,25	4,25	4,25	4,5	4	4,125	4,125
Q8	2,5	2,25	2,25	2,375	2,125	2,5	2,375	2,25	2,25	3	2,875	2,375
Q9	4,125	4,5	3,5	4	4,5	4,25	3,875	4,25	4,75	4,125	4,25	4,5
Q10	2,5	2,25	2,25	2,25	2,25	2,5	2,5	2,375	2,375	3,625	3,125	2,875

TABLE A.3: Moyennes des mesures pour tous les participants.

B

Documents de l'évaluation de l'audience virtuelle

Cette annexe regroupe les documents et questionnaires utilisés pour l'étude visant à évaluer le système d'entrainement à la prise de parole en public décrit au chapitre 8.

Les six premiers documents (section B.1) correspodent aux indications données aux participants de l'étude. Ces documents étaient mis à disposition des participants lors de la phase d'entrainement (*i.e.* deuxième et troisième présentations).

Ensuite, les deux documents suivants (section B.2) sont les résumés des présentations proposés aux participants pour leur préparation avant l'étude.

Enfin, les sept derniers documents sont les questionnaires remplis par les participants lors de l'étude. Les quatre premiers étaient remplis avant l'étude (section B.3), les trois suivants après l'étude (section B.4). La table B.1 rapporte les moyennes et écart-types des réponses des participants au questionnaire d'auto-évaluation.

B.1 INDICATIONS AUX PARTICIPANTS

Gaze training

How a public speaker looks at his audience is very important. Here are a few quotes from toastmasters:

"Don't just pass your gaze throughout the room; try to focus on individual listeners and create a bond with them by looking them directly in the eyes for five to 10 seconds."
http://www.toastmasters.org

"Maintain eye contact with your audience. Make friends with the group in the center, but remember everyone in the room."
http://www.addisonsinglestoastmasters.com/

This scenario will help you train how you look at an audience during public speaking.

- Try to look at individual characters in the audience for a few seconds.
- While there are more listeners on the front screen audience, you should not forget the left screen audience.
- Try to avoid looking at your slides or away for too long.

Gaze training – Gauges

How a public speaker looks at his audience is very important. Here are a few quotes from toastmasters:

"Don't just pass your gaze throughout the room; try to focus on individual listeners and create a bond with them by looking them directly in the eyes for five to 10 seconds."
http://www.toastmasters.org

"Maintain eye contact with your audience. Make friends with the group in the center, but remember everyone in the room."
http://www.addisonsinglestoastmasters.com/

This scenario will help you train how you look at an audience during public speaking.

- Try to look at individual characters in the audience for a few seconds.
- While there are more listeners on the front screen audience, you should not forget the left screen audience.
- Try to avoid looking at your slides or away for too long.

There will be a gauge for the left and right part of the audience.

If you haven't looked at a part of the screen for too long, the corresponding gauge will turn red.

If you have looked at them recently, it will stay green.

Gaze training – Interactive audience

How a public speaker looks at his audience is very important. Here are a few quotes from toastmasters:

"Don't just pass your gaze throughout the room; try to focus on individual listeners and create a bond with them by looking them directly in the eyes for five to 10 seconds."
http://www.toastmasters.org

"Maintain eye contact with your audience. Make friends with the group in the center, but remember everyone in the room."
http://www.addisonsinglestoastmasters.com/

This scenario will help you train how you look at an audience during public speaking.

- Try to look at individual characters in the audience for a few seconds.
- While there are more listeners on the front screen audience, you should not forget the left screen audience.
- Try to avoid looking at your slides or away for too long.

If a character in the audience hasn't been looked at for too long, they might try to get your attention by clearing their throat or coughing.

Speech training

It is important, when speaking in public, to not use too many filler words, such as *"err" or "um"*.

"Most beginning speakers are afraid of pauses. They believe their audience will think they are inarticulate if they pause to think of what to say next, so they use filler words to avoid the silence."

"Filler words are insidious because they are invisible to the speaker, but not to the listener. To help members become aware of this verbal clutter, Toastmasters clubs designate an Ah-Counter, who tracks filler words used by all speakers during a meeting and then discloses the results at the end."
(Toastmasters magazine, February 2011, "Cutting out filler words")

This scenario will help you reduce the amount of filler words you use during public speaking.

- Try to think about your sentence before saying it.
- Don't be afraid to pause.
- The audience will not interrupt you.

Speech training – Gauges

It is important, when speaking in public, to not use too many filler words, such as *“err” or “um”*.

“Most beginning speakers are afraid of pauses. They believe their audience will think they are inarticulate if they pause to think of what to say next, so they use filler words to avoid the silence.”

“Filler words are insidious because they are invisible to the speaker, but not to the listener. To help members become aware of this verbal clutter, Toastmasters clubs designate an Ah-Counter, who tracks filler words used by all speakers during a meeting and then discloses the results at the end.”
(Toastmasters magazine, February 2011, “Cutting out filler words”)

This scenario will help you reduce the amount of filler words you use during public speaking.

- Try to think about your sentence before saying it.
- Don’t be afraid to pause.
- The audience will not interrupt you.

In this scenario, there will be a gauge at the top of the screen.

If it is fully green, you have used no filler words in the last ten seconds.

If there is a red section, it represents the amount of filler words you used in the last ten seconds.

Speech training – Interactive audience

It is important, when speaking in public, to not use too many filler words, such as *"err" or "um"*.

"Most beginning speakers are afraid of pauses. They believe their audience will think they are inarticulate if they pause to think of what to say next, so they use filler words to avoid the silence."

"Filler words are insidious because they are invisible to the speaker, but not to the listener. To help members become aware of this verbal clutter, Toastmasters clubs designate an Ah-Counter, who tracks filler words used by all speakers during a meeting and then discloses the results at the end."
(Toastmasters magazine, February 2011, "Cutting out filler words")

This scenario will help you reduce the amount of filler words you use during public speaking.

- Try to think about your sentence before saying it.
- Don't be afraid to pause.
- The audience will not interrupt you.

In this scenario, characters in the audience will react to your filler words.

- If you speak without filler words, they may nod and may tend to lean towards you.
- If you use filler words, some characters may shake their heads and lean backwards.

B.2 PROPOSITIONS DE RÉSUMÉS DES PRÉSENTATIONS

Dear participant,

Your task is to present to the audience, the city of Los Angeles.
Here are some facts about the city:

When Los Angeles was founded in 1781, 44 people (14 families) lived in El Pueblo de Nuestra Senora la Reina de Los Angeles de la Porciuncula (Town of Our Lady the Queen of the Angeles of the Small Portion). The population grew, but the name shrank to simply "Los Angeles."

Now, Los Angeles is the largest city in California and the second-largest urban area in the nation: the Los Angeles five-county area has a population of almost 20 million.

Los Angeles has a Subtropical-Mediterranean climate. The average annual temperature in downtown is 75° F during the day and 57° F at night.

Los Angeles is billed as the "Creative capital of the world". Iconic landmarks of the entertainment industry include the Hollywood sign, the Hollywood walk of fame, and Grauman's Chinese Theater.

Rodeo drive, pictured in *Pretty Woman*, a three-block stretch of designer shops, is great for people-watching.

Los Angeles offers a lot for entertainment. The Universal Studios Hollywood is a theme park that features a tour of some of the company's most famous stages. Disneyland is an amazing venue for family entertainment.

Sports fan can also enjoy the famous Dodgers Stadium housing the city's baseball team. The Staples Center is a multi-purpose venue, home to the Los Angeles Lakers, the Los Angeles Clippers, and the Los Angeles Kings.

There are 841 museums and art galleries in Los Angeles County. In fact, Los Angeles has more museums per capita than any other city in the world. Some of the notable museums are the Getty Center and the Museum of Contemporary Art.

The Walt Disney Concert Hall, centerpiece of the Music Center, is home to the prestigious Los Angeles Philharmonic.

Los Angeles is also famous for its beach cities. Venice is known for its canals, beaches, and the Ocean Front Walk, a two-and-a-half-mile pedestrian-only promenade that features performers, fortune-tellers, artists, and vendors.

The famous Santa Monica pier houses a small amusement Park, Pacific Park, and is the official end of route 66.

Many Los Angeles beaches are world-renowned surfer spots such as Malibu Surfrider's beach, which was the first ever World Surfing Reserve

Dear participant,

Your task is to present to the audience, the new Lancôme VISIONNAIRE Advanced Skin Corrector product Here are some facts about this product:

VISIONNAIRE is an all-in-one product. It works against the appearance of wrinkles*, pores* and skin texture imperfections and can be applied to the face up to the eye contour area, morning and evening.
*Clinical study

Whatever her age, if a woman wants to correct the appearance of skin texture imperfections such as visible pores, unevenness, acne marks, fine lines or wrinkles, then VISIONNAIRE is ideal. It has been tested on women from 38 to 60 years old.

VISIONNAIRE can be used all year round, every day morning and evening. It is safe for all skin types, even sensitive skins.

During our consumer studies, VISIONNAIRE was applied twice a day, morning and night, for 6 consecutive weeks.

At 8 weeks*:

- Overall skin appearance is improved:
- Wrinkles look reduced, smoothed out:
- Pores look less visible:
- Skin imperfections look minimized:
- Skin texture is refined:

*Based upon the self-assessment of the total of all consumers tested at 4 and 8 weeks

VISIONNAIRE works against the appearance of wrinkles*, but also improves the appearance of enlarged pores* and skin texture imperfections, such as acne marks. In addition, VISIONNAIRE is safe for all skin types, even sensitive skins. It can be applied every day, morning and night under your Lancôme moisturizer.
*Clinical study

The molecule in VISIONNAIRE, LR 2412, is light-stable and does not create photo-sensibility. So, VISIONNAIRE can be used all year round, even in the summer. However, in addition to Lancôme Visionnaire, we always recommend to protect the skin against the harmful effects of UV rays, particularly, of course, in the summer.

VISIONNAIRE has been tested on all skin types, even sensitive, and was well tolerated throughout all the tests we performed.

VISIONNAIRE has been tested on all skin tones** with excellent results.
**Multi-ethnic study -test conducted on the 6 Fitzpatrick phototypes

VISIONNAIRE corrects the appearance of wrinkles*-pores*-unevenness. Like all serums, it should be used morning and night under your favourite Lancôme moisturizer. It can be integrated into your routine in addition to your usual products. It has a non-oily formula which means that it will not make your skin sticky if you layer it with other products.
*Clinical study

The women recruited to this study originally considered undergoing a cosmetic procedure such as hyaluronic acid injections, chemical peeling or laser sessions*.
*After 4 weeks of consumer use. Consumer evaluations of women 35 to 49 years tempted by hyaluronic acid, laser or chemical peeling. Results not equal to a medical procedure

In the case you do not have wrinkels, VISIONNAIRE will help perfect your skin's texture and visibly correct other issues that tarnish the beauty of your skin, such as enlarged pores* or texture imperfections, such as acne marks.
*Clinical study

VISIONNAIRE is suitable for all skin types, even sensitive skins and can be used all year round, morning and night even in the summer time.

B.3 QUESTIONNAIRES AVANT L'INTERACTION

Demographic Questionnaire

I. General Information

Participant Number_____

Age_____

Gender_____

Education:
____Elementary school
____Middle school
____High school
____Bachelor's degree
____Master's degree
____Doctoral degree
____Other (please specify):_____

Ethnicity:
____African American
____Asian
____Hispanic
____Caucasian (white
____Native America
____Other (please specify):_____

Do you wear glasses? Y N

II. Computer Experience
1. Do you own a personal computer? Y N
2. Do you use a computer at your job? Y N
3. On average how many hours a week do you spend on a computer? ____
1) 0-5 hrs
2) 5-10 hrs
3) 10-20 hrs
4) 20-40 hrs
5) 40+

4. How many computer courses have you taken?_____

5. How would you rate your computer competency using the scale below?___
1) Completely inexperienced
2) Inexperienced
3) Average
4) Experienced
5) Very experienced

6. Indicate the average number of hours a week you do the following computer activities:
a. Word processing ____
b. Programming ____
c. Games playing ____
d. Data entry/processing _____
e. Graphic design/ art _____
f. Surfing the internet ____
g. Emailing ____
h. Other ____ ____

Participant #:

	1 Disagree strongly	2 Disagree a little	3 Neither agree nor disagree	4 Agree a little	5 Agree strongly
I see myself as someone who…					
…is reserved					
…is generally trusting					
…tends to be lazy					
…is relaxed, handles stress well					
…has few artistic interests					
…is outgoing, sociable					
…tends to find fault with others					
…does a thorough job					
…gets nervous easily					
…has an active imagination					

Personal Report of Confidence as a Public Speaker (PRCS)

1 I look forward to an opportunity to speak in public.
2 My hands tremble when I try to handle objects on the platform.
3 I am in constant fear of forgetting my speech.
4 Audiences seem friendly when I address them.
5 While preparing a speech I am in a constant state of anxiety.

6 At the conclusion of a speech I feel I have had a pleasant experience.
7 I dislike to use my body and voice expressively.
8 My thoughts become confused and jumbled when I speak before an audience.
9 I have no fear of facing an audience.
10 Although I am nervous just before getting up I soon forget my fears and enjoy the experience.
11 I face the prospect of making a speech with complete confidence.
12 I feel that I am in complete possession of myself while speaking.
13 I prefer ti have notes on the platform in case I forget my speech.
14 I like to observe the reactions of my audience to my speech.
15 Although I talk fluently with friends, I am at a loss for words on the platform.
16 I feel relaxed and comfortable while speaking.

17 Although I do not enjoy speaking in public I do not particularly dread it.
18 I always avoid speaking in public if possible.
19 The faces of my audience are blurred when I look at them.

20 I feel disgusted with myself after trying to address a group of people.
21 I enjoy preparing a talk.
22 My mind is clear when I face an audience.
23 I am fairly fluent.
24 I perspire and tremble just before getting up.
25 My posture feels strained and unnatural.
26 I am fearful and tensed all the while I am speaking before a group of people.
27 I find the prospect of speaking mildly pleasant.
28 It is difficult for me to calmly search my mind for the right words to express my thoughts.
29 I am terrified at the thought of speaking before a group of people.
30 I have a feeling of alertness in facing an audience.

Participant #

Please imagine what you have typically felt and thought to yourself during any kind of public speaking situation.

Imagining these situations, how much do you agree with the statements given below?

Please rate the degree of your agreement on a scale between 0 and 4.

		0 Disagree strongly	1 Disagree a little	2 Neither agree nor disagree	3 Agree a little	4 Agree strongly
1	What do I have to lose it's worth a try					
2	I'm a loser					
3	This is an awkward situation but I can handle it					
4	A failure in this situation would be more proof of my incapacity					
5	Even if thing's don't go well, it's no catastrophe					
6	I can handle everything					
7	What I say will probably sound stupid					
8	I'll probably “bomb out” away					
9	Instead of worrying I could concentrate on what I want to say					
10	I feel awkward and dumb; they're bound to notice					

B.4 QUESTIONNAIRES APRÈS L'INTERACTION

Participant #:

Please take a moment to consider the interaction you just had and give us your feedback.

This scale consists of a number of words that describe different feelings and emotions.

Please indicate the degree to which you felt each of the following during the interaction, by marking with an "x" the desired option. Please consider each of them separately.

	1 Never	2 Rarely	3 Every once in a while	4 Often	5 Always
Upset					
Hostile					
Alert					
Ashamed					
Inspired					
Nervous					
Determined					
Attentive					
Afraid					
Active					

Self-Assessment

Participant #:

The following questions only relate to the presentation you just gave in front of the virtual audience.
Self score by inserting "x" in the relevant score box for each statement.

	How successful do you think you were in:	Very weak 1	2	3	4	Very Strong 5
1	Hiding your emotions throughout the presentation?					
2	Maintaining a fitting speaking rate in the presentation?					
2	Maintaining a proper eye-contact with the audience?					
3	Maintaining a fitting speech loudness in the presentation?					
4	Giving the talk with adequate intonation?					
5	Communicating the idea to the audience?					
6	Avoiding hesitations in speech?					
7	Emphasizing the presentation's contents by gesturing when appropriate?					
8	Avoiding gesturing too much?					
9	Controlling the stutter?					
10	Maintaining a logical flow of the presentation?					
11	Respecting the presentation's structure in your speech?					

A score of 1 indicates that you consider your performance to be very weak in that criterion.
A score of 5 indicates that you consider your performance to be very strong in that criterion.s

On a scale from 0 (not at all nervous) to 100 (extremely nervous) please indicate:

Overall, how nervous were you during the presentation?
Overall, how successful was your presentation?

Participant #:

Please answer the following questions by marking the relevant number. In particular, remember that these questions are asking you about how you felt at the end of the presentation.

1 To what extent did the virtual audience hold your attention?
Not at all 1 2 3 4 5 A lot

2 To what extent did you feel you were focused on the virtual audience?
Not at all 1 2 3 4 5 A lot

3 How much effort did you out into giving the presentation?
Very little 1 2 3 4 5 A lot

4 Did you feel that you were trying your best?
Not at all 1 2 3 4 5 Very much so

5 To what extent did you lose track of time?
Not at all 1 2 3 4 5 A lot

6 To what extent did you feel consciously aware of presenting in front of a virtual audience whilst talking?
Not at all 1 2 3 4 5 Very much so

7 To what extent did you forget about your everyday concerns?
Not at all 1 2 3 4 5 A lot

8 To what extent were you aware of yourself in your surroundings?
Not at all 1 2 3 4 5 Very aware

9 To what extent did you notice events taking place around you?
Not at all 1 2 3 4 5 A lot

10? Did you feel the urge at any point to stop playing and see what was happening around you?
Not at all 1 2 3 4 5 Very much so

11 To what extent did you feel that you were interacting with the VA/environment?
Not at all 1 2 3 4 5 Very much so

12 To what extent did you feel as though you were separated from your real-world environment?
Not at all 1 2 3 4 5 Very much so

13 To what extent did you feel the game was something you were experiencing, rather than something you were just doing?
Not at all 1 2 3 4 5 Very much so

14 To what extent was your sense of being in the VA environment stronger than your sense of being in the real world?

Not at all 1 2 3 4 5 Very much so

15 At any point did you find yourself so involved that you were unaware you were presenting to a virtual audience?
Not at all 1 2 3 4 5 Very much so

16 To what extent did you feel as though you were moving in front of the virtual audience according to your own will?
Not at all 1 2 3 4 5 Very much so

17 To what extent did you find presenting in front of a virtual audience challenging?
Not at all 1 2 3 4 5 Very difficult

18 To what extent did you feel motivated while presenting?
Not at all 1 2 3 4 5 A lot

19 To what extent did you find presenting in front of a virtual audience easy?
Not at all 1 2 3 4 5 Very much so

20 To what extent did you feel you were making progress towards the end of the presentation?
Not at all 1 2 3 4 5 A lot

21 How well do you think you performed during the presentation?
Very poor 1 2 3 4 5 Very well

22 To what extent did you feel emotionally attached to the virtual audience?
Not at all 1 2 3 4 5 Very much so

23 How much did you want to "win"?
Not at all 1 2 3 4 5 Very much so

24 Were you in suspense about whether or not you would give a good or bad presentation in front of the virtual audience?
Not at all 1 2 3 4 5 Very much so

25 At any point did you find yourself become so involved that you wanted to speak to the virtual audience directly?
Not at all 1 2 3 4 5 Very much so

26 To what extent did you enjoy the graphics and the imagery?
Not at all 1 2 3 4 5 A lot

27 How much would you say you enjoyed giving the presentation?
Not at all 1 2 3 4 5 A lot

28 Would you like to repeat this experience?
Definitely not 1 2 3 4 5 Definitely yes

29 How useful do you think this tool can help you with improving your public speaking skills?
Not at all 1 2 3 4 5 A lot

30 Would you train with this tool to improve your public speaking skills, if given the chance?

Definitely not 1 2 3 4 5 Definitely yes

How immersed did you feel? (10 – very immersed; 0 – not at all immersed)

If you have any other suggestions or ideas about the experience, please use this space to express your thoughts:

SA_{Qi}	Texte	CTRL		RD		AVI	
		μ	σ	μ	σ	μ	σ
Q1	To what extent did the virtual audience hold your attention ?	3,44	1,09	3,6	1,4	4,53	0,52
Q2	To what extent did you feel you were focused on the virtual audience ?	3,69	1,25	3,8	1,21	4,2	1,15
Q3	How much effort did you out into giving the presentation ?	4,56	0,63	4,6	0,63	4,53	0,64
Q4	Did you feel that you were trying your best ?	4,44	1,03	4,53	0,74	4,73	0,59
Q5	To what extent did you lose track of time ?	3,38	1,67	3,27	1,1	3,67	1,11
Q6	To what extent did you feel consciously aware of presenting in front of a virtual audience whilst talking ?	2,63	1,2	3,47	1,25	3,53	1,68
Q7	To what extent did you forget about your everyday concerns ?	4,5	0,89	4,6	0,63	4,6	0,51
Q8	To what extent were you aware of yourself in your surroundings ?	4,19	0,98	3,87	1,19	3,8	1,47
Q9	To what extent did you notice events taking place around you ?	2,5	1,41	2,8	1,61	2,53	1,55
Q10	Did you feel the urge at any point to stop playing and see what was happening around you ?	1,94	1,39	1,73	0,96	1,33	0,62
Q11	To what extent did you feel that you were interacting with the VA/environment ?	3,25	1,29	3,4	1,59	3,47	1,41
Q12	To what extent did you feel as though you were separated from your real-world environment ?	3,56	1,15	3,4	1,45	3,47	1,19
Q13	To what extent did you feel the game was something you were experiencing, rather than something you were just doing ?	3,25	1	3,27	1,39	3,6	1,18
Q14	To what extent was your sense of being in the VA environment stronger than your sense of being in the real world ?	3	1,46	3	1,13	3,2	1,21
Q15	At any point did you find yourself so involved that you were unaware you were presenting to a virtual audience ?	3,5	1,26	3,2	1,15	3,07	1,58
Q16	To what extent did you feel as though you were moving in front of the virtual audience according to your own will ?	3,94	0,77	4,4	0,74	4,2	1,15
Q17	To what extent did you find presenting in front of a virtual audience challenging ?	2	1,03	2,8	1,47	2,93	1,16
Q18	To what extent did you feel motivated while presenting ?	4,19	0,75	4,33	0,9	4,33	0,98
Q19	To what extent did you find presenting in front of a virtual audience easy ?	4	0,97	3,87	1,13	4,27	0,88
Q20	To what extent did you feel you were making progress towards the end of the presentation ?	4,44	0,89	4,2	0,77	4,73	0,59
Q21	How well do you think you performed during the presentation ?	3,69	1,08	4	0,93	4,27	0,88
Q22	To what extent did you feel emotionally attached to the virtual audience ?	2,56	1,15	2,6	1,3	2,73	1,16
Q23	How much did you want to “win” ?	4,06	1,44	4,4	0,83	4,33	1,11
Q24	Were you in suspense about whether or not you would give a good or bad presentation in front of the virtual audience ?	3,25	1,34	4,13	1,19	3,47	1,3
Q25	At any point did you find yourself become so involved that you wanted to speak to the virtual audience directly ?	3,06	1,29	3,53	1,51	2,93	1,49
Q26	To what extent did you enjoy the graphics and the imagery ?	3,56	1,21	3,4	1,5	3,6	1,06
Q27	How much would you say you enjoyed giving the presentation ?	4,31	0,79	4,2	0,94	4,53	1,06
Q28	Would you like to repeat this experience ?	4,44	0,96	4,73	0,8	4,87	0,35
Q29	How useful do you think this tool can help you with improving your public speaking skills ?	4,75	0,58	4,73	0,59	4,93	0,26
Q30	Would you train with this tool to improve your public speaking skills, if given the chance ?	4,69	0,7	4,73	0,59	4,87	0,35
Q31	How immersed did you feel ? (10 – very immersed ; 0 – not at all immersed)	7,93	2,81	7,85	2,15	7	2,65

TABLE B.1: Réponses des participants aux questionnaires d'auto-évaluation. CTRL correspond à la condition de contrôle (audience passive), RD à la condition de retours directs, et AVI à l'audience virtuelle interactive.

Liste des publications

Publications dans des conférences internationales

M. Chollet, T. Wörtwein, L.-P. Morency, A. Shapiro, S. Scherer. *Exploring Feedback Learning Strategies to Improve Public Speaking: An Interactive Virtual Audience Framework*. Pervasive and Ubiquitous Computing (UbiComp'15), Septembre 2015, Osaka, Japon.

A. Ben Youssef, **M. Chollet**, H. Jones, N. Sabouret, C. Pelachaud, M. Ochs. *Towards a Socially Adaptive Virtual Agent*. Intelligent Virtual Agents (IVA'15), Août 2015, Delft, Pays-Bas.

M. Chollet, M. Ochs, C. Pelachaud. *From Non-verbal Signals Sequence Mining to Bayesian Networks for Interpersonal Attitudes Expression*. Intelligent Virtual Agents (IVA'14), pp 120-133, Août 2014, Boston, États-Unis. **Nomination pour le prix du meilleur article**.

M. Chollet, M. Ochs, C. Pelachaud. *Mining a multimodal corpus for non-verbal signals sequences conveying attitudes*. Language Resources and Evaluation Conference (LREC'14), pp 3417-3424, Mai 2014, Reykjavik, Islande.

H. Jones, **M. Chollet**, M. Ochs, C. Pelachaud, N. Sabouret. *Expressing social attitudes in virtual agents for social coaching*. Autonomous Agents and Multi-Agent Systems (AAMAS'14), pp 1409-1410, Mai 2014, Paris, France.

M. Chollet, G. Sratou, A. Shapiro, L.-P. Morency and S. Scherer. *An Interactive Virtual Audience Platform for Public Speaking Training - Demonstration*. Autonomous Agents and Multi-Agent Systems (AAMAS'14), pp 1657-1658, Mai 2014, Paris, France.

K. Anderson, E. André, T. Baur, S. Bernardini, **M. Chollet**, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard, H. Jones, M. Ochs, C. Pelachaud, K. Porayska-Pomsta, P. Rizzo, N. Sabouret. *The TARDIS framework: intelligent virtual agents for social coaching in job interviews*. Advances in Computer Entertainment (ACE'13), pp 476-491, Novembre 2013, Enschede, Pays-Bas.

M. Chollet, M. Ochs, C. Clavel, C. Pelachaud. *A multimodal corpus approach to the design of virtual recruiters*. Affective Computing and Intelligent Interaction (ACII'13), pp 19-24, 2013, Genève, Suisse.

Publications dans des conférences nationales

M. Ochs, Y. Ding, N. Fourati, **M. Chollet**, B. Ravenet, F. Pecune, N. Glas, K. Prépin, C. Clavel et C. Pelachaud. *Vers des Agents Conversationnels Animés Socio-Affectifs*. Interaction Humain-Machine (IHM'13), pp 69-78, Novembre 2013, Bordeaux, France. **Nomination pour le prix du meilleur article**.

Publications dans des ateliers internationaux

A. Ben Youssef, **M. Chollet**, H. Jones, N. Sabouret, C. Pelachaud, M. Ochs. *An Architecture for a Socially Adaptive Virtual Recruiter in Job Interview Simulations*. 3rd International Workshop on Intelligent Digital Games for Empowerment and Inclusion (IDGEI'15), Mars 2015, Atlanta, USA.

F. Pécune, A. Cafaro, **M. Chollet**, P. Philippe, C. Pelachaud. *Suggestions for Extending SAIBA with the VIB Platform*. IVA'14 Workshop on Architectures and Standards for Intelligent Virtual Agents (WASIVA'14), Août 2014, Boston, États-Unis.

N. Sabouret, H. Jones, M. Ochs, **M. Chollet**, C. Pelachaud. *Expressing social attitudes in virtual agents for social training games*. 2nd International Workshop on Intelligent Digital Games for Empowerment and Inclusion (IDGEI'14), Février 2014, Haifa, Israël.

M. Chollet, M. Ochs, C. Pelachaud. *Investigating non-verbal behaviors conveying interpersonal stances*. European Symposium on Multimodal Communication (MMSym'13), Octobre 2013, La Vallette, Malte.

M. Chollet, M. Ochs, C. Pelachaud. *A multimodal corpus for the study of non-verbal behavior conveying interpersonal stance*. IVA'13 Workshop Multimodal Corpora: Beyond Audio and Video (MMC'13), Septembre 2013, Édimbourg, Royaume-Uni.

Publications dans des ateliers nationaux

H. Jones, **M. Chollet**, M. Ochs, N. Sabouret, C. Pelachaud. *Expressing social attitudes in virtual agents for social coaching*. Workshop Affect, Compagnon Artificiel, Interaction, Juin 2014, Rouen, France.

M. Chollet, M. Ochs, C. Pelachaud. *Interpersonal Stance Recognition Using Non-Verbal Signals on Several Time Windows*. Workshop Affect, Compagnon Artificiel, Interaction, Novembre 2012, Grenoble, France.

Bibliographie

[Agrawal & Srikant, 1994] Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases* (pp. 487–499). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 94

[Allwood, 2008] Allwood, J. (2008). Multimodal corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook* (pp. 207–225). Berlin: Mouton de Gruyter. 61, 64

[Allwood et al., 2007a] Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., & Paggio, P. (2007a). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3-4), 273–287. 64

[Allwood et al., 2007b] Allwood, J., Kopp, S., Grammer, K., Ahlsén, E., Oberzaucher, E., & Koppensteiner, M. (2007b). The analysis of embodied communicative feedback in multimodal corpora: a prerequisite for behavior simulation. *Language Resources and Evaluation*, 41(3-4), 255–272. 27, 64, 84, 186

[Allwood et al., 1992] Allwood, J., Nivre, J., & Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback. *Journal of semantics*, 9(1), 1–26. 32

[Anderson et al., 2013] Anderson, K., André, E., Baur, T., Bernardini, S., Chollet, M., Chryssafidou, E., Damian, I., Ennis, C., Egges, A., Gebhard, P., Jones, H., Ochs, M., Pelachaud, C., Porayska-Pomsta, K., Rizzo, P., & Sabouret, N. (2013). The TARDIS framework: Intelligent virtual agents for social coaching in job interviews. In D. Reidsma, H. Katayose, & A. Nijholt (Eds.), *Advances in Computer Entertainment*, volume 8253 of *Lecture Notes in Computer Science* (pp. 476–491). Springer International Publishing. 149

[Antunes & Oliveira, 2001] Antunes, C. M. & Oliveira, A. L. (2001). Temporal data mining: An overview. In *Workshop on Temporal Data Mining in the ACM International Conference on Knowledge Discovery and Data Mining* (pp. 1–13). 181

[Argyle, 1988] Argyle, M. (1988). *Bodily Communication*. University paperbacks. Methuen. 8, 18, 22, 24, 28, 30, 31, 50, 178

[Argyle & Cook, 1976] Argyle, M. & Cook, M. (1976). *Gaze and Mutual Gaze*. Cambridge University Press. 28, 32, 46, 73

[Aronoff et al., 1988] Aronoff, J., Barclay, A. M., & Stevenson, L. A. (1988). The recognition of threatening facial stimuli. *Journal of personality and social psychology*, 54(4), 647–655. 30

[Aylett & Pidcock, 2007] Aylett, M. & Pidcock, C. (2007). The CereVoice characterful speech synthesiser SDK. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, & D. Pelé (Eds.), *Intelligent Virtual Agents*, volume 4722 of *Lecture Notes in Computer Science* (pp. 413–414). Springer Berlin Heidelberg. 115

[Aylett et al., 2014] Aylett, R., Hall, L., Tazzyman, S., Endrass, B., André, E., Ritter, C., Nazir, A., Paiva, A., Höfstede, G., & Kappas, A. (2014). Werewolves, cheats, and cultural sensitivity. In *Proceedings of the 13th International Conference on Autonomous Agents and Multi-Agent Systems* (pp. 1085–1092). International Foundation for Autonomous Agents and Multiagent Systems. 39

[Aylett et al., 2007] Aylett, R., Vala, M., Sequeira, P., & Paiva, A. (2007). Fearnot!–an emergent narrative approach to virtual dramas for anti-bullying education. In *Virtual Storytelling. Using Virtual Reality Technologies for Storytelling* (pp. 202–205). Springer Berlin Heidelberg. 39

[Bakeman & Brownlee, 1980] Bakeman, R. & Brownlee, J. R. (1980). The strategic use of parallel play: A sequential analysis. *Child Development*, 51(3), 873–878. 84

[Bakeman & Quera, 2011] Bakeman, R. & Quera, V. (2011). *Sequential Analysis and Observational Methods for the Behavioral Sciences*. Cambridge University Press. 84

[Bales, 1950] Bales, R. (1950). *A Set of Categories for the Analysis of Small Group Interaction.Channels of Communication in Small Groups*. Bobbs-Merrill. 74

[Ballin et al., 2004] Ballin, D., Gillies, M., & Crabtree, I. (2004). A framework for interpersonal attitude and non-verbal communication in improvisational visual

media production. In *Proceedings of the 1st European Conference on Visual Media Production*. 25, 45, 46, 52

[Batrinca et al., 2013] Batrinca, L., Stratou, G., Shapiro, A., Morency, L.-P., & Scherer, S. (2013). Cicero - towards a multimodal virtual audience platform for public speaking training. In R. Aylett, B. Krenn, C. Pelachaud, & H. Shimodaira (Eds.), *Intelligent Virtual Agents*, volume 8108 of *Lecture Notes in Computer Science* (pp. 116–128). Springer Berlin Heidelberg. 6, 148, 150, 154, 155, 161

[Bee et al., 2009] Bee, N., Franke, S., & Andrea, E. (2009). Relations between facial display, eye gaze and head tilt: Dominance perception variations of virtual agents. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction* (pp. 1–7). IEEE. 48, 52

[Bee et al., 2010] Bee, N., Pollock, C., André, E., & Walker, M. (2010). Bossy or Wimpy: Expressing Social Dominance by Combining Gaze and Linguistic Behaviors. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Intelligent Virtual Agents*, volume 6356 of *Lecture Notes in Computer Science* chapter 28, (pp. 265–271). Springer Berlin Heidelberg. 48, 52, 127, 144

[Bernardini et al., 2012] Bernardini, S., Porayska-Pomsta, K., Smith, T., & Avramides, K. (2012). Building autonomous social partners for autistic children. In Y. Nakano, M. Neff, A. Paiva, & M. Walker (Eds.), *Intelligent Virtual Agents*, volume 7502 of *Lecture Notes in Computer Science* (pp. 46–52). Springer Berlin Heidelberg. 3, 39

[Bevacqua et al., 2012] Bevacqua, E., De Sevin, E., Hyniewska, S. J., & Pelachaud, C. (2012). A listener model: introducing personality traits. *Journal on Multimodal User Interfaces*, 6(1-2), 27–38. 126

[Bevacqua et al., 2010] Bevacqua, E., Pammi, S., Hyniewska, S., Schröder, M., & Pelachaud, C. (2010). Multimodal backchannels for embodied conversational agents. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Intelligent Virtual Agents*, volume 6356 of *Lecture Notes in Computer Science* (pp. 194–200). Springer Berlin Heidelberg. 126

[Bickmore & Picard, 2005] Bickmore, T. W. & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions in Computer-Human Interaction*, 12(2), 293–327. 47, 52

[Blache et al., 2010] Blache, P., Bertrand, R., Bigi, B., Bruno, E., Cela, E., Espesser, R., Ferré, G., Guardiola, M., Hirst, D., Magro, E.-P., Martin, J.-C., Meunier, C., Morel, M.-A., Murisasco, E., Nesterenko, I., Nocera, P., Pallaud, B., Prévot, L., Priego-Valverde, B., Seinturier, J., Tan, N., Tellier, M., & Rauzy, S. (2010). Multimodal annotation of conversational data. In *Proceedings of the 4th Linguistic Annotation Workshop* (pp. 186–191). Association for Computational Linguistics. 68, 70

[Boersma & Weenink, 2001] Boersma, P. & Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9-10), 341–345. 72

[Bosseler & Massaro, 2003] Bosseler, A. & Massaro, D. W. (2003). Development and evaluation of a computer-animated tutor for vocabulary and language learning in children with autism. *Journal of autism and developmental disorders*, 33(6), 653–672. 39

[Briton & Hall, 1995] Briton, N. J. & Hall, J. A. (1995). Beliefs about female and male nonverbal communication. *Sex Roles*, 32(1-2), 79–90. 28

[Brown, 1987] Brown, P. (1987). *Politeness: Some universals in language usage*. Cambridge University Press. 128

[Brundage et al., 2006] Brundage, S. B., Graap, K., Gibbons, K. F., Ferrer, M., & Brooks, J. (2006). Frequency of stuttering during challenging and supportive virtual reality job interviews. *Journal of fluency disorders*, 31(4), 325–339. 43

[Brunet et al., 2009] Brunet, P., Donnan, H., McKeown, G., Douglas-Cowie, E., & Cowie, R. (2009). Social signal processing: What are the relevant variables? and in what ways do they relate? In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops* (pp. 1–6). IEEE. 73

[Burgoon & Dunbar, 2006] Burgoon, J. & Dunbar, N. (2006). *Nonverbal Expressions of Dominance and Power in Human Relationships*, (pp. 279–299). SAGE Publications, Inc. 32

[Burgoon et al., 1984] Burgoon, J. K., Buller, D. B., Hale, J. L., & de Turck, M. A. (1984). Relational Messages Associated with Nonverbal Behaviors. *Human Communication Research*, 10(3), 351–378. 8, 27, 28, 31, 50

[Burgoon & Hale, 1984] Burgoon, J. K. & Hale, J. L. (1984). The fundamental topoi of relational communication. *Communication Monographs*, 51(3), 193–214. 24

[Burgoon & Le Poire, 1999] Burgoon, J. K. & Le Poire, B. A. (1999). Nonverbal cues and interpersonal judgments: Participant and observer perceptions of intimacy, dominance, composure, and formality. *Communication Monographs*, 66(2), 105–124. 8, 27, 28, 29, 31, 52

[Cafaro, 2014] Cafaro, A. (2014). *First Impressions in Human-Agent Virtual Encounters*. PhD thesis, Reykjavik University, Iceland. 5, 25, 52, 59

[Cafaro et al., 2012] Cafaro, A., Vilhjálmsson, H., Bickmore, T., Heylen, D., Jóhannsdóttir, K., & Valgarðsson, G. (2012). First impressions: Users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In Y. Nakano, M. Neff, A. Paiva, & M. Walker (Eds.), *Intelligent Virtual Agents*, volume 7502 of *Lecture Notes in Computer Science* (pp. 67–80). Springer Berlin Heidelberg. 49

[Cafaro et al., 2014] Cafaro, A., Vilhjálmsson, H., Bickmore, T., Heylen, D., & Pelachaud, C. (2014). Representing communicative functions in saiba with a unified function markup language. In T. Bickmore, S. Marsella, & C. Sidner (Eds.), *Intelligent Virtual Agents*, volume 8637 of *Lecture Notes in Computer Science* (pp. 81–94). Springer International Publishing. 104

[Calbris, 2011] Calbris, G. (2011). *Elements of Meaning in Gesture*, volume 5 of *Gesture studies*. John Benjamins Publishing Company. 54

[Callejas et al., 2014] Callejas, Z., Ravenet, B., Ochs, M., & Pelachaud, C. (2014). A computational model of social attitudes for a virtual recruiter. In *Proceedings of the 13th International Conference on Autonomous Agents and Multi-Agent Systems* (pp. 93–100). International Foundation for Autonomous Agents and Multiagent Systems. 50, 127, 128, 130, 144

[Carletta et al., 2006] Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., & Wellner, P. (2006). The AMI meeting corpus: A pre-announcement. In S. Renals & S. Bengio

(Eds.), *Proceedings of the 2nd International Conference on Machine Learning for Multimodal Interaction* (pp. 28–39). Springer-Verlag, Berlin, Heidelberg. 68, 70

[Carney et al., 2005] Carney, D. R., Hall, J. A., & LeBeau, L. S. (2005). Beliefs about the nonverbal expression of social power. *Journal of Nonverbal Behavior*, 29(2), 105–123. 8, 27, 28, 29, 30, 75

[Cassell, 2007] Cassell, J. (2007). Body language: Lessons from the near-human. In J. Riskin (Ed.), *Genesis Redux: Essays in the History and Philosophy of Artificial Life* (pp. 346–374). University of Chicago Press. 61, 62

[Cassell et al., 2001] Cassell, J., Nakano, Y. I., Bickmore, T. W., Sidner, C. L., & Rich, C. (2001). Non-verbal cues for discourse structure. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (pp. 114–123). Association for Computational Linguistics. 73

[Cassell et al., 2000] Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (2000). *Embodied Conversational Agents*. MIT Press. 1

[Cassell & Thorisson, 1999] Cassell, J. & Thorisson, K. R. (1999). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(4-5), 519–538. 45, 73

[Cassell et al., 2004] Cassell, J., Vilhjálmsson, H. H., & Bickmore, T. (2004). Beat: the behavior expression animation toolkit. In *Life-Like Characters* (pp. 163–185). Springer. 5, 52

[Charniak et al., 1993] Charniak, E., Hendrickson, C., Jacobson, N., & Perkowitz, M. (1993). Equations for part-of-speech tagging. In *Proceedings of the 11th National Conference on Artificial Intelligence* (pp. 784–789). Menlo Park: AAAI Press/MIT Press. 86

[Chindamo et al., 2012] Chindamo, M., Allwood, J., & Ahlsén, E. (2012). Some suggestions for the study of stance in communication. In *Proceedings of the 4th ASE/IEEE International Conference on Social Computing* (pp. 617–622). IEEE. 22, 23

[Condon & Osgton, 1971] Condon, W. S. & Osgton, W. D. (1971). Speech and body motion synchrony of the speaker-hearer. In D. Horton & J. Jenkins (Eds.), *The perception of language* (pp. 150–184). New York: Academic Press. 107

[Cook & Smith, 1975] Cook, M. & Smith, J. (1975). The role of gaze in impression formation. *British journal of social and clinical psychology*, 14(1), 19–25. 46

[Cortes & Gatti, 1965] Cortes, J. B. & Gatti, F. M. (1965). Physique and self-description of temperament. *Journal of Consulting Psychology*, 29(5), 432. 26

[Cosnier & Vaysse, 1997] Cosnier, J. & Vaysse, J. (1997). Sémiotique des gestes communicatifs. *Nouveaux actes sémiotiques 52-53-54*, (pp. 7–28). 16, 22

[Costa et al., 2001] Costa, M., Menzani, M., & Bitti, P. E. R. (2001). Head canting in paintings: An historical study. *Journal of Nonverbal Behavior*, 25(1), 63–73. 29

[Courgeon et al., 2014] Courgeon, M., Céline, C., & Martin, J.-C. (2014). Modeling facial signs of appraisal during interaction: impact on users' perception and behavior. In *Proceedings of the 13th annual conference on Autonomous Agents and Multi-Agent Systems* (pp. 765–772). International Foundation for Autonomous Agents and Multiagent Systems. 55, 56, 57

[Cowie et al., 2011] Cowie, R., Douglas-Cowie, E., McRorie, M., Sneddon, I., Devillers, L., & Amir, N. (2011). Issues in data collection. In P. Petta, C. Pelachaud, & R. Cowie (Eds.), *Emotion-Oriented Systems* (pp. 197–212). Springer-Verlag, Berlin, Heidelberg. 63

[Cowie et al., 2000] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). 'FEELTRACE': An instrument for recording perceived emotion in real time. In R. Cowie, E. Douglas-Cowie, & M. Schroede (Eds.), *Speech and Emotion: Proceedings of the ISCA Workshop*. Newcastle, Co. Down. 66

[Cowie & McKeown, 2010] Cowie, R. & McKeown, G. (2010). Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme, SEMAINE D6b deliverable. available from "downloads" at http://www.semaine-project.eu/. 67, 78, 79

[Cowie et al., 2012] Cowie, R., McKeown, G., & Douglas-Cowie, E. (2012). Tracing emotion: An overview. *International Journal of Synthetic Emotions*, 3(1), 1–17. 66, 67, 78

[Culpeper et al., 2003] Culpeper, J., Bousfield, D., & Wichmann, A. (2003). Impoliteness revisited: with special reference to dynamic and prosodic aspects. *Journal of Pragmatics*, 35(10), 1545–1579. 128

[Darwin, 1872] Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. Harper Perennial. 30

[Debras & Cienki, 2012] Debras, C. & Cienki, A. (2012). Some uses of head tilts and shoulder shrugs during human interaction, and their relation to stancetaking. In *Proceedings of the 4th ASE/IEEE International Conference on Social Computing* (pp. 932–937). IEEE Computer Society. 29

[Deng & Neumann, 2008] Deng, Z. & Neumann, U. (2008). *Data-Driven 3D Facial Animation*. Springer. 5

[DeVault et al., 2014] DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., & Morency, L.-P. (2014). Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 13th International Conference on Autonomous Agents and Multi-agent Systems* (pp. 1061–1068). International Foundation for Autonomous Agents and Multiagent Systems. 152

[DiMatteo et al., 1986] DiMatteo, M. R., Hays, R. D., & Prince, L. M. (1986). Relationship of physicians' nonverbal communication skill to patient satisfaction, appointment noncompliance, and physician workload. *Health Psychology*, 5(6), 581–594. 148

[Ding et al., 2013] Ding, Y., Pelachaud, C., & Artières, T. (2013). Modeling multimodal behaviors from speech prosody. In R. Aylett, B. Krenn, C. Pelachaud, & H. Shimodaira (Eds.), *Intelligent Virtual Agents*, volume 8108 of *Lecture Notes in Computer Science* (pp. 217–228). Springer Berlin Heidelberg. 53

[Ding et al., 2014] Ding, Y., Prepin, K., Huang, J., Pelachaud, C., & Artières, T. (2014). Laughter animation synthesis. In *Proceedings of the 13th International Conference on Autonomous Agents and Multi-Agent Systems* (pp. 773–780). International Foundation for Autonomous Agents and Multiagent Systems. 5, 53

[Douglas-Cowie et al., 2003] Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech communication*, 40(1), 33–60. 63

[Douglas-Cowie et al., 2007] Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., Martin, J.-C., Devillers, L., Abrilian, S., Batliner, A., Amir, N., & Karpouzis, K. (2007). The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. In A. Paiva, R. Prada, & R. Picard (Eds.), *Affective Computing and Intelligent Interaction*, volume 4738 of *Lecture Notes in Computer Science* (pp. 488–500). Springer Berlin Heidelberg. 63

[Du Bois, 2007] Du Bois, J. W. (2007). The stance triangle. In R. Englebretson (Ed.), *Stancetaking in discourse: Subjectivity, evaluation, interaction* (pp. 139–182). John Benjamins Amsterdam, The Netherlands, and Philadelphia, PA. 22

[Dunbar & Burgoon, 2005] Dunbar, N. E. & Burgoon, J. K. (2005). Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships*, 22(2), 207–233. 27, 28, 30, 32

[Dybkjær & Bernsen, 2004] Dybkjær, L. & Bernsen, N. O. (2004). Recommendations for natural interactivity and multimodal annotation schemes. In J.-C. Martin, L. B. Kühnlein, P. Paggio, & R. Catizone (Eds.), *Proceedings of the 3rd Workshop on Multimodal Corpora at LREC'04* (pp. 5–8). 64

[Ekman, 1972] Ekman, P. (1972). Universals and cultural differences in facial expression of emotion. In R. Colre (Ed.), *Nebraska Symposium on Motivation*, volume 19 (pp. 207–283). University of Nebraska Press, Lincoln. 57

[Ekman, 2007] Ekman, P. (2007). The directed facial action task. In *Handbook of emotion elicitation and assessment* (pp. 47–53). Oxford University Press Oxford, UK. 5

[Ekman & Friesen, 1976] Ekman, P. & Friesen, V. (1976). *Pictures of Facial Affect*. Palo Alto: Consulting Psychologists Press. 5

[Ekman & Friesen, 1977] Ekman, P. & Friesen, V. (1977). *Manual for the Facial Action Coding System*. Palo Alto: Consulting Psychologists Press. 56, 57, 64, 75

[Ekman & Friesen, 1969] Ekman, P. & Friesen, W. (1969). The repertoire of nonverbal behavior: Categories, origins, usage and coding. *Semiotica*, 1(1), 49–98. 14, 19, 22, 30

[el Kaliouby & Robinson, 2005] el Kaliouby, R. & Robinson, P. (2005). Real-time inference of complex mental states from facial expressions and head gestures. In B. Kisačanin, V. Pavlović, & T. Huang (Eds.), *Real-Time Vision for Human-Computer Interaction* (pp. 181–200). Springer US. 55

[Ferreira & Azevedo, 2005] Ferreira, P. G. & Azevedo, P. J. (2005). Protein sequence classification through relevant sequence mining and bayes classifiers. In C. Bento, A. Cardoso, & G. Dias (Eds.), *Progress in Artificial Intelligence*, volume 3808 of *Lecture Notes in Computer Science* (pp. 236–247). Springer Berlin Heidelberg. 88

[Foster & Oberlander, 2007] Foster, M. E. & Oberlander, J. (2007). Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Language Resources and Evaluation*, 41(3-4), 305–323. 61

[Fournier-Viger et al., 2014] Fournier-Viger, P., Gomariz, A., Gueniche, T., Soltani, A., Wu., C., & Tseng, V. S. (2014). SPMF: a Java Open-Source Pattern Mining Library. *Journal of Machine Learning Research*, 15, 3389–3393. 94

[Fukayama et al., 2002] Fukayama, A., Ohno, T., Mukawa, N., Sawaki, M., & Hagita, N. (2002). Messages embedded in gaze of interface agents - impression management with agent's gaze. In *Proceedings of the 2002 SIGCHI conference on Human factors in computing systems* (pp. 41–48). New York, New York, USA: ACM Press. 46, 52, 139

[Gatica-Perez, 2009] Gatica-Perez, D. (2009). Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing*, 27(12), 1775–1787. 27

[Gebhard et al., 2014] Gebhard, P., Baur, T., Damian, I., Mehlmann, G., Wagner, J., & André, E. (2014). Exploring interaction strategies for virtual characters to induce stress in simulated job interviews. In *Proceedings of the 13th International Conference on Autonomous Agents and Multi-agent Systems* (pp. 661–668). International Foundation for Autonomous Agents and Multiagent Systems. 126, 127

[Gebhard et al., 2012] Gebhard, P., Mehlmann, G., & Kipp, M. (2012). Visual scenemaker—a tool for authoring interactive virtual characters. *Journal on Multimodal User Interfaces*, 6(1-2), 3–11. 122

[Gifford, 1991] Gifford, R. (1991). Mapping nonverbal behavior on the interpersonal circle. *Journal of Personality and Social Psychology*, 61(2), 279. 29

[Gifford, 1994] Gifford, R. (1994). A lens-mapping framework for understanding the encoding and decoding of interpersonal dispositions in nonverbal behavior. *Journal of Personality and Social Psychology*, 66(2), 398. 8, 27, 28, 29

[Gifford & Hine, 1994] Gifford, R. & Hine, D. W. (1994). The role of verbal behavior in the encoding and decoding of interpersonal dispositions. *Journal of Research in Personality*, 28(2), 115–132. 8, 27

[Giraud et al., 2013] Giraud, T., Soury, M., Hua, J., Delaborde, A., Tahon, M., Gomez Jauregui, D., Eyharabide, V., Filaire, E., Le Scanff, C., Devillers, L., Isableu, B., & Martin, J. (2013). Multimodal expressions of stress during a public speaking task: Collection, annotation and global analyses. In *Proceedings of the 5th Humaine Association Conference Affective Computing and Intelligent Interaction* (pp. 417–422). IEEE Computer Society. 68, 70

[Glass et al., 1982] Glass, C. R., Merluzzi, T. V., Biever, J. L., & Larsen, K. H. (1982). Cognitive assessment of social anxiety: Development and validation of a self-statement questionnaire. *Cognitive Therapy and Research*, 6(1), 37–55. 27

[Graesser et al., 2003] Graesser, A. C., Moreno, K., Marineau, J., Adcock, A., Olney, A., & Person, N. (2003). Autotutor improves deep learning of computer literacy: Is it the dialog or the talking head? In U. Hoppe, M. Verdejo, & J. Kay (Eds.), *Artificial Intelligence in Education: Shaping the Future of Learning Through Intelligent Technologies*, volume 97 of *Frontiers in artificial intelligence and applications* (pp. 47–54). IOS Press. 1

[Gratch et al., 2014] Gratch, J., Lucas, G. M., King, A., & Morency, L.-P. (2014). It's only a computer: The impact of human-agent interaction in clinical interviews. In *Proceedings of the 13th annual conference on Autonomous Agents and Multi-Agent Systems* (pp. 85–92). International Foundation for Autonomous Agents and Multiagent Systems. 39

[Gratch & Marsella, 2005] Gratch, J. & Marsella, S. (2005). Lessons from emotion psychology for the design of lifelike characters. *Applied Artificial Intelligence*, 19(3-4), 215–233. 2

[Gratch et al., 2006] Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R., & Morency, L.-P. (2006). Virtual rapport. In J. Gratch, M. Young, R. Aylett, D. Ballin, & P. Olivier (Eds.), *Intelligent Virtual Agents*, volume 4133 of *Lecture Notes in Computer Science* (pp. 14–27). Springer Berlin Heidelberg. 39, 125

[Gregor & Richmond, 2010] Gregor, H. & Richmond, K. (2010). Comparison of hmm and tmdn methods for lip synchronisation. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech)* (pp. 454–457). International Speech Communication Association. 52

[Grillon et al., 2006] Grillon, H., Riquier, F., Herbelin, B., & Thalmann, D. (2006). Virtual reality as a therapeutic tool in the confines of social anxiety disorder treatment. *International journal on disability and human development*, 5(3), 243–250. 3, 39, 40, 41, 148

[Guillame-Bert & Crowley, 2012] Guillame-Bert, M. & Crowley, J. L. (2012). Learning temporal association rules on symbolic time sequences. In *Proceedings of the 4th Asian Conference on Machine Learning* (pp. 159–174). 181, 186

[Gunes & Piccardi, 2006] Gunes, H. & Piccardi, M. (2006). Creating and annotating affect databases from face and body display: A contemporary survey. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics* (pp. 2426–2433). 63, 68

[Hadar et al., 1983] Hadar, U., Steiner, T., Grant, E., & Rose, F. C. (1983). Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2(1), 35–46. 29

[Hall, 1966] Hall, E. (1966). *Distances in man: The hidden dimension*. Double Day. 50

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Exploration Newsletter*, 11(1), 10–18. 111

[Harrigan et al., 1985] Harrigan, J., Oxman, T., & Rosenthal, R. (1985). Rapport expressed through nonverbal behavior. *Journal of Nonverbal Behavior*, 9(2), 95–110. 29

[Harris et al., 2002] Harris, S. R., Kemmerling, R. L., & North, M. M. (2002). Brief virtual reality therapy for public speaking anxiety. *Cyberpsychology and Behavior*, 5, 543–550. 39, 40, 41, 148

[Hart et al., 2013] Hart, J., Gratch, J., & Marsella, S. C. (2013). How virtual reality training can win friends and influence people. In C. Best, G. Galanis, J. Kerry, & R. Sottilare (Eds.), *Fundamental Issues in Defense Training and Simulation*, Human Factors in Defense (pp. 235–249). Ashgate. 38, 39, 147, 149, 154, 170

[Hartholt et al., 2013] Hartholt, A., Traum, D., Marsella, S. C., Shapiro, A., Stratou, G., Leuski, A., Morency, L.-P., & Gratch, J. (2013). All together now: Introducing the virtual human toolkit. In R. Aylett, B. Krenn, C. Pelachaud, & H. Shimodaira (Eds.), *Intelligent Virtual Agents*, volume 8108 of *Lecture Notes in Computer Science* (pp. 368–381). Springer Berlin Heidelberg. 5, 52

[Hartmann et al., 2006] Hartmann, B., Mancini, M., & Pelachaud, C. (2006). Implementing expressive gesture synthesis for embodied conversational agents. In S. Gibet, N. Courty, & J.-F. Kamp (Eds.), *Gesture in Human-Computer Interaction and Simulation*, volume 3881 of *Lecture Notes in Computer Science* (pp. 188–199). Springer Berlin Heidelberg. 129, 131

[Herbelin, 2005] Herbelin, B. (2005). *Virtual reality exposure therapy for social phobia*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne. 41

[Hess et al., 2007] Hess, U., Adams Jr, R. B., & Kleck, R. E. (2007). Looking at you or looking elsewhere: The influence of head orientation on the signal value of emotional facial expressions. *Motivation and Emotion*, 31(2), 137–144. 5

[Hess et al., 2000] Hess, U., Blairy, S., & Kleck, R. E. (2000). The influence of facial emotion displays, gender, and ethnicity on judgments of dominance and affiliation. *Journal of Nonverbal behavior*, 24(4), 265–283. 30

[Heylen, 2008] Heylen, D. (2008). Listening heads. In I. Wachsmuth & G. Knoblich (Eds.), *Modeling Communication with Robots and Virtual Humans*, volume 4930

of *Lecture Notes in Computer Science* (pp. 241–259). Springer Berlin Heidelberg. 29

[Heylen et al., 2008] Heylen, D., Kopp, S., Marsella, S., Pelachaud, C., & Vilhjálmsson, H. (2008). The next step towards a function markup language. In H. Prendinger, J. Lester, & M. Ishizuka (Eds.), *Intelligent Virtual Agents*, volume 5208 of *Lecture Notes in Computer Science* (pp. 270–280). Springer Berlin Heidelberg. 9, 54, 104, 107, 179

[Hirschberg, 1993] Hirschberg, J. (1993). Pitch accent in context predicting intonational prominence from text. *Artificial Intelligence*, 63(1), 305–340. 26

[Hirschberg & Grosz, 1992] Hirschberg, J. & Grosz, B. (1992). Intonational features of local and global discourse structure. In *Proceedings of the Workshop on Speech and Natural Language* (pp. 441–446). Stroudsburg, PA, USA: Association for Computational Linguistics. 26

[Hofmann & DiBartolo, 2000] Hofmann, S. G. & DiBartolo, P. M. (2000). An instrument to assess self-statements during public speaking: Scale development and preliminary psychometric properties. *Behavior Therapy*, 31(3), 499–515. 160

[Hook et al., 2008] Hook, J. N., Smith, C. A., & Valentiner, D. P. (2008). A short-form of the personal report of confidence as a speaker. *Personality and Individual Differences*, 44(6), 1306–1313. 160

[Hoque & Picard, 2014] Hoque, M. & Picard, R. (2014). Rich nonverbal sensing technology for automated social skills training. *Computer*, 47(4), 28–35. 3

[Hoque et al., 2013] Hoque, M. E., Courgeon, M., Martin, J.-C., Mutlu, B., & Picard, R. W. (2013). MACH: My automated conversation coach. In *Proceedings of the 15th International Conference on Pervasive and Ubiquitous Computing* (pp. 697–706). New York, NY, USA: ACM. 43, 44, 45, 149

[Hua et al., 2014] Hua, J., Le Scanff, C., Larue, J., José, F., Martin, J.-C., Devillers, L., & Filaire, E. (2014). Global stress response during a social stress test: impact of alexithymia and its subfactors. *Psychoneuroendocrinology*, 50, 53–61. 69

[Hunyadi et al., 2012] Hunyadi, L., Szekrenyes, I., Borbely, A., & Kiss, H. (2012). Annotation of spoken syntax in relation to prosody and multimodal pragmatics.

In *3rd International Conference on Cognitive Infocommunications* (pp. 537–541). IEEE. 70, 181

[Jack et al., 2014] Jack, R. E., Garrod, O. G., & Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology*, 24(2), 187–192. 6, 32, 52, 83, 177

[Jaillet et al., 2006] Jaillet, S., Laurent, A., & Teisseire, M. (2006). Sequential patterns for text categorization. *Intelligent Data Analysis*, 10(3), 199–214. 113

[Jennett et al., 2008] Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., & Walton, A. (2008). Measuring and defining the experience of immersion in games. *International journal of human-computer studies*, 66(9), 641–661. 161

[Johnsen et al., 2007] Johnsen, K., Raij, A., Stevens, A., Lind, D. S., & Lok, B. (2007). The validity of a virtual human experience for interpersonal skills education. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 1049–1058). ACM. 39

[Johnson et al., 2000] Johnson, W. L., Rickel, J. W., & Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, 11(1), 47–78. 1, 149

[Johnson & Valente, 2009] Johnson, W. L. & Valente, A. (2009). Tactical language and culture training systems: using AI to teach foreign languages and cultures. *AI Magazine*, 30(2), 72. 39

[Jones et al., 2014] Jones, H., Chollet, M., Ochs, M., Sabouret, N., & Pelachaud, C. (2014). Expressing social attitudes in virtual agents for social coaching. In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems* (pp. 1409–1410). International Foundation for Autonomous Agents and Multiagent Systems. 123

[Jonsson, 2011] Jonsson, G. (2011). *Hidden Temporal Patterns In Interaction*. PhD thesis, University of Aberdeen. 85

[Jonsson et al., 2010] Jonsson, G. K., Anguera, M., Sánchez-Algarra, P., Olivera, C., Campanico, J., Castañer, M., Torrents, C., Dinusova, M., & Chaverri, J.

(2010). Application of t-pattern detection and analysis in sports research. *Open Sports Sciences Journal*, 3, 95–104. 85

[Keating et al., 1981] Keating, C. F., Mazur, A., Segall, M. H., Cysneiros, P. G., Kilbride, J. E., Leahy, P., Divale, W. T., Komin, S., Thurman, B., & Wirsing, R. (1981). Culture and the perception of social dominance from facial expression. *Journal of personality and social psychology*, 40(4), 615. 30

[Keltner, 1995] Keltner, D. (1995). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*, 68, 441–454. 6, 31, 52, 83, 88, 177

[Keltner & Haidt, 1999] Keltner, D. & Haidt, J. (1999). Social functions of emotions at four levels of analysis. *Cognition & Emotion*, 13(5), 505–521. 27

[Kendon, 1967] Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta psychologica*, 26(1), 22–63. 5, 28, 46, 47

[Kendon, 1983] Kendon, A. (1983). Gesture and speech. how they interact. In J. M. Wiemann & R. P. Harrison (Eds.), *Nonverbal Interaction* (pp. 13–45). London: Sage. 19

[Kendon, 1988] Kendon, A. (1988). How gestures can become like words. In F. Poyatos (Ed.), *Cross-cultural perspectives in Nonverbal Communication* (pp. 131–141). Hogrefe. 20

[Kendon, 1990] Kendon, A. (1990). *Conducting interaction: Patterns of behavior in focused encounters*, volume 7 of *Studies in Interactional Sociolinguistics*. Cambridge University Press, New York. 49

[Kim & Baylor, 2006] Kim, Y. & Baylor, A. L. (2006). A social-cognitive framework for pedagogical agents as learning companions. *Educational Technology Research and Development*, 54(6), 569–596. 1

[Kim et al., 2008] Kim, Y., Schmidt, E., & Emelle, L. (2008). Moodswings: A collaborative game for music mood label collection. In *Proceedings of the 9th International Conference on Music Information Retrieval* (pp. 231–236). Philadelphia, USA. 66

[Kipp, 2012] Kipp, M. (2012). *Multimedia Annotation, Querying, and Analysis in Anvil*, (pp. 351–367). John Wiley & Sons, Inc. 65

[Kirschbaum et al., 1993] Kirschbaum, C., Pirke, K.-M., & Hellhammer, D. H. (1993). The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2), 76–81. 68

[Kita et al., 1998] Kita, S., van Gijn, I., & van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.), *Gesture and Sign Language in Human-Computer Interaction*, volume 1371 of *Lecture Notes in Computer Science* (pp. 23–35). Springer Berlin Heidelberg. 21

[Knapp et al., 2013] Knapp, M., Hall, J., & Horgan, T. (2013). *Nonverbal communication in human interaction*. Cengage Learning. 1, 5, 25, 27

[Knight, 2011] Knight, D. (2011). The future of multimodal corpora. *Revista Brasileira de Linguística Aplicada*, 11(2), 391–415. 63, 68

[Knutson, 1996] Knutson, B. (1996). Facial expressions of emotion influence interpersonal trait inferences. *Journal of Nonverbal Behavior*, 20(3), 165–182. 30

[Kopp et al., 2006] Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Thórisson, K., & Vilhjálmsson, H. (2006). Towards a common framework for multimodal generation: The behavior markup language. In J. Gratch, M. Young, R. Aylett, D. Ballin, & P. Olivier (Eds.), *Intelligent Virtual Agents*, volume 4133 of *Lecture Notes in Computer Science* (pp. 205–217). Springer Berlin Heidelberg. 104

[Krumhuber et al., 2007] Krumhuber, E., Manstead, A., & Kappas, A. (2007). Temporal aspects of facial displays in person and expression perception: The effects of smile dynamics, head-tilt, and gender. *Journal of Nonverbal Behavior*, 31(1), 39–56. 30

[Kwon et al., 2013] Kwon, J. H., Powell, J., & Chalmers, A. (2013). How level of realism influences anxiety in virtual reality environments for a job interview. *International Journal of Human-Computer Studies*, 71(10), 978–987. 43

[Labov, 1978] Labov, W. (1978). *Sociolinguistic Patterns*, volume 4 of *Conduct and Communication*. University of Pennsylvania Press. 63

[LaFrance, 1982] LaFrance, M. (1982). Posture mirroring and rapport. In M. Davis (Ed.), *Interaction Rhythms: Periodicity in Communicative Behavior* (pp. 279–299). New York: Human Sciences Press. 28, 32, 119

[Lane et al., 2013] Lane, H. C., Hays, M. J., Core, M. G., & Auerbach, . (2013). Learning intercultural communication skills with virtual humans: Feedback and fidelity. *Journal of Educational Psychology*, 105(4), 1026–1035. 3, 39, 40, 149

[Lane & Wray, 2012] Lane, H. C. & Wray, R. E. (2012). Individualized cultural and social skills learning with virtual humans. In P. J. Durlach & A. M. Lesgold (Eds.), *Adaptive Technologies for Training and Education* (pp. 204–221). New York: Cambridge University Press. 2

[Leary, 1957] Leary, T. (1957). *Interpersonal diagnosis of personality*. New York: Ronald. 24

[Lee & Marsella, 2006] Lee, J. & Marsella, S. (2006). Nonverbal behavior generator for embodied conversational agents. In J. Gratch, M. Young, R. Aylett, D. Ballin, & P. Olivier (Eds.), *Intelligent Virtual Agents*, volume 4133 of *Lecture Notes in Computer Science* (pp. 243–255). Springer Berlin Heidelberg. 5, 52

[Lee & Marsella, 2011] Lee, J. & Marsella, S. (2011). Modeling side participants and bystanders: The importance of being a laugh track. In H. Vilhjálmsson, S. Kopp, S. Marsella, & K. Thórisson (Eds.), *Intelligent Virtual Agents*, volume 6895 of *Lecture Notes in Computer Science* (pp. 240–247). Springer Berlin Heidelberg. 25, 48, 52, 59

[Lee & Marsella, 2010] Lee, J. & Marsella, S. C. (2010). Predicting speaker head nods and the effects of affective information. *IEEE Transactions on Multimedia*, 12(6), 552–562. 53, 58, 86

[Linssen et al., 2014] Linssen, J., Theune, M., & Heylen, D. (2014). Taking things at face value: How stance informs politeness of virtual agents. In M. Conci, V. Dignum, M. Funk, & D. Heylen (Eds.), *Proceedings of the Workshop on Computers As Social Actors*, volume 1119 of *CEUR Workshop Proceedings* (pp. 71–82). 128

[MacIntyre et al., 1997] MacIntyre, P. D., Thivierge, K. A., & MacDonald, J. R. (1997). The effects of audience interest, responsiveness, and evaluation on public

speaking anxiety and related variables. *Communication Research Reports*, 14(2), 157–168. 149

[Magnusson, 2000] Magnusson, M. S. (2000). Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior Research Methods, Instruments, Computers*, 32, 93–110. 85

[Mairesse & Walker, 2011] Mairesse, F. & Walker, M. A. (2011). Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3), 455–488. 127

[Malatesta et al., 2007] Malatesta, L., Caridakis, G., Raouzaiou, A., & Karpouzis, K. (2007). Agent personality traits in virtual environments based on appraisal theory predictions. *Artificial and Ambient Intelligence, Language, Speech and Gesture for Expressive Characters*, 7, 1621–1630. 55, 56

[Mancini & Pelachaud, 2007] Mancini, M. & Pelachaud, C. (2007). Dynamic behavior qualifiers for conversational agents. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, & D. Pelé (Eds.), *Intelligent Virtual Agents*, volume 4722 of *Lecture Notes in Computer Science* (pp. 112–124). Springer Berlin Heidelberg. 52, 183

[Mancini & Pelachaud, 2008] Mancini, M. & Pelachaud, C. (2008). The FML - APML language. In *Proceedings of the First FML workshop, AAMAS'08*. Estoril, Portugal. 107, 115

[Marioryad & Busso, 2013] Marioryad, S. & Busso, C. (2013). Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations. In *Proceedings of the 5th Humaine Association Conference Affective Computing and Intelligent Interaction* (pp. 85–90). IEEE Computer Society. 67

[Marsella et al., 2013] Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., & Shapiro, A. (2013). Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (pp. 25–35). New York, NY, USA: ACM. 54

[Martin et al., 2006] Martin, J.-C., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M., & Pelachaud, C. (2006). Du corpus vidéo à l'agent expressif : Utilisation

des différents niveaux de représentation multimodale et émotionnelle. *Revue d'intelligence artificielle*, 20(4-5), 477–498. 5

[Martínez & Yannakakis, 2011] Martínez, H. P. & Yannakakis, G. N. (2011). Mining multimodal sequential patterns: a case study on affect detection. In *Proceedings of the 13th International Conference on Multimodal Interfaces* (pp. 3–10). New York, NY, USA: ACM. 88

[Mason et al., 2005] Mason, M. F., Tatkow, E. P., & Macrae, C. N. (2005). The look of love gaze shifts and person perception. *Psychological Science*, 16(3), 236–239. 32

[Masseglia et al., 2004] Masseglia, F., Teisseire, M., & Poncelet, P. (2004). Extraction de motifs séquentiels. problèmes et méthodes. *Ingénierie des Systèmes d'Information*, 9(3-4), 183–210. 87

[Mast, 2002] Mast, M. S. (2002). Dominance as expressed and inferred through speaking time. *Human Communication Research*, 28(3), 420–450. 27

[McClave, 2000] McClave, E. Z. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32(7), 855–878. 29

[McCroskey, 1970] McCroskey, J. C. (1970). *Measures of communication-bound anxiety*. Taylor & Francis. 160

[McKeown et al., 2010] McKeown, G., Valstar, M. F., Cowie, R., & Pantic, M. (2010). The semaine corpus of emotionally coloured character interactions. In *Proceedings of the 2010 IEEE International Conference on Multimedia and Expo* (pp. 1079–1084). IEEE Computer Society. 68, 70

[McNeill, 1992] McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Psychology/cognitive science. University of Chicago Press. 19, 20, 22, 27, 88

[McNeill, 2005] McNeill, D. (2005). *Gesture and thought*. University of Chicago Press. 20

[Mehrabian, 1969] Mehrabian, A. (1969). Significance of posture and position in the communication of attitude and status relationships. *Psychological Bulletin*, 71(5), 359. 25

[Mehrabian, 1977] Mehrabian, A. (1977). *Nonverbal communication*. Transaction Publishers. 28

[Mehrabian, 1981] Mehrabian, A. (1981). *Silent Messages: Implicit Communication of Emotions and Attitudes*. Wadsworth Publishing Company. 13

[Metallinou & Narayanan, 2013] Metallinou, A. & Narayanan, S. (2013). Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *Proceedings of the 10th IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 1–8). IEEE. 78

[Mignault & Chaudhuri, 2003] Mignault, A. & Chaudhuri, A. (2003). The many faces of a neutral face: Head tilt and perception of dominance and emotion. *Journal of Nonverbal Behavior*, 27(2), 111–132. 29

[Milde & Gut, 2002] Milde, J. & Gut, U. (2002). The tasx-environment: an xml-based toolset for time aligned speech corpora. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (pp. 1922–1927). European Language Resources Association. 65

[Miller et al., 2014] Miller, C., Quek, F., & Morency, L.-P. (2014). : (pp. 273–280).: ACM Press. 99

[Moreno et al., 2001] Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction*, 19(2), 177–213. 39, 154

[Munhall et al., 2004] Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility head movement improves auditory speech perception. *Psychological science*, 15(2), 133–137. 29

[Nagel et al., 2007] Nagel, F., Kopiez, R., Grewe, O., & Altenmüller, E. (2007). Emujoy: Software for continuous measurement of perceived emotions in music. *Behavior Research Methods*, 39(2), 283–290. 66

[Naumann et al., 2009] Naumann, L. P., Vazire, S., Rentfrow, P. J., & Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality and Social Psychology Bulletin*, 35, 1661–1671. 26

[Niewiadomski et al., 2011] Niewiadomski, R., Hyniewska, S. J., & Pelachaud, C. (2011). Constraint-based model for synthesis of multimodal sequential expressions of emotions. *IEEE Transaction on Affective Computing*, 2(3), 134–146. 5, 56, 58, 59

[Noller, 1985] Noller, P. (1985). Video primacy—a further look. *Journal of Nonverbal Behavior*, 9(1), 28–47. 144

[Ogan & Lane, 2010] Ogan, A. & Lane, H. C. (2010). Virtual learning environments for culture and intercultural competence. In E. Blanchard & D. Allard (Eds.), *Handbook of research on culturally-aware information technology: Perspectives and models* (pp. 501–519). Hershey, PA: Information Science Reference. 39

[Otta et al., 1994] Otta, E., Lira, B. B. P., Delevati, N. M., Cesar, O. P., & Pires, C. S. G. (1994). The effect of smiling and of head tilting on person perception. *The Journal of psychology*, 128(3), 323–331. 29, 30

[Paleari et al., 2007] Paleari, M., Grizard, A., & Lisetti, C. L. (2007). Adapting psychologically grounded facial emotional expressions to different anthropomorphic embodiment platforms. In D. Wilson & G. Sutcliffe (Eds.), *Proceedings of the 20th International Florida Artificial Intelligence Research Society Conference (FLAIRS)* (pp. 565–570). AAAI Press. 55, 56

[Pan et al., 2007] Pan, X., Gillies, M., Sezgin, T. M., & Loscos, C. (2007). Expressing complex mental states through facial expressions. In A. Paiva, R. Prada, & R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction*, volume 4738 of *Lecture Notes in Computer Science* (pp. 745–746). Springer. 55, 56, 58

[Pandzic & Forchheimer, 2002] Pandzic, I. S. & Forchheimer, R. (2002). *MPEG-4 facial animation*. John Wiley & Sons. 5

[Park et al., 2012] Park, S., Gratch, J., & Morency, L.-P. (2012). I already know your answer: Using nonverbal behaviors to predict immediate outcomes in a dyadic negotiation. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (pp. 19–22). ACM. 149

[Paul, 1966] Paul, G. L. (1966). *Insight vs. Desensitization in Psychotherapy: An Experiment in Anxiety Reduction*. Stanford University Press. 42

[Pelachaud et al., 1996] Pelachaud, C., Badler, N. I., & Steedman, M. (1996). Generating facial expressions for speech. *Cognitive science*, 20(1), 1–46. 5

[Pertaub et al., 2002] Pertaub, D.-P., Slater, M., & Barker, C. (2002). An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators and virtual environments*, 11(1), 68–78. 3, 40, 41, 42, 148

[Peterson, 2005] Peterson, R. T. (2005). An examination of the relative effectiveness of training in nonverbal communication: Personal selling implications. *Journal of Marketing Education*, 27(2), 143–150. 149

[Pez et al., 2014] Pez, A.-M., Philippe, P., Donval, B., & Pelachaud, C. (2014). Virtual Interactive Behavior : une architecture modulaire pour ACA. In G. Bailly, M. Ochs, & A. Pauchet (Eds.), *Actes du Workshop Affects, Compagnons Artificiels, Interaction* (pp. 91–92). 131

[Picard, 1997] Picard, R. W. (1997). *Affective computing*. MIT press. 63

[Poggi, 2003] Poggi, I. (2003). Mind markers. In M. R. N. Trigo & I. Poggi (Eds.), *Gestures. Meaning and use.* University Fernando Pessoa Press, Oporto. 5, 14, 17, 22, 65, 105, 107, 109

[Poggi, 2007] Poggi, I. (2007). *Mind, Hands, Face and Body: A Goal and Belief View of Multimodal Communication*. Weidler Buchverlag Berlin. 31, 32

[Poggi & Pelachaud, 1998] Poggi, I. & Pelachaud, C. (1998). Performative faces. *Speech Communication*, 26(1–2), 5–21. 29

[Poggi et al., 2005] Poggi, I., Pelachaud, C., de Rosis, F., Carofiglio, V., & De Carolis, B. (2005). Greta. a believable embodied conversational agent. In O. Stock & M. Zancanaro (Eds.), *Multimodal Intelligent Information Presentation*, volume 27 of *Text, Speech and Language Technology* (pp. 3–25). Springer Netherlands. 5, 52

[Rabiner, 1990] Rabiner, L. (1990). A tutorial on hidden Markov models and selected applications in speech recognition. In A. Waibel & K.-F. Lee (Eds.), *Readings in Speech Recognition* (pp. 267–296). Kaufmann, San Mateo, CA. 86

[Rammstedt & John, 2007] Rammstedt, B. & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1), 203–212. 160

[Ravenet et al., 2013] Ravenet, B., Ochs, M., & Pelachaud, C. (2013). From a user-created corpus of virtual agent's non-verbal behavior to a computational model of interpersonal attitudes. In R. Aylett, B. Krenn, C. Pelachaud, & H. Shimodaira (Eds.), *Intelligent Virtual Agents*, volume 8108 of *Lecture Notes in Computer Science* (pp. 263–274). Springer Berlin Heidelberg. 25, 50, 51, 52, 129

[Rehm & André, 2008] Rehm, M. & André, E. (2008). From annotated multimodal corpora to simulated human-like behaviors. In I. Wachsmuth & G. Knoblich (Eds.), *Modeling Communication with Robots and Virtual Humans* (pp. 1–17). Springer. 62

[Richmond & McCroskey, 2000] Richmond, V. & McCroskey, J. (2000). *Nonverbal Behavior in Interpersonal Relations*. Allyn and Bacon. 28, 50

[Rizzo et al., 2011] Rizzo, A., Kenny, P., & Parsons, D. T. (2011). Intelligent virtual patients for training clinical skills. *Journal of Virtual Reality and Broadcasting*, 8(3). 39

[Rosenberg & Hirschberg, 2005] Rosenberg, A. & Hirschberg, J. (2005). Acoustic/prosodic and lexical correlates of charismatic speech. In *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH)* (pp. 513–516). ISCA. 161

[Rowe et al., 2010] Rowe, J., Shores, L., Mott, B., & Lester, J. (2010). Integrating learning and engagement in narrative-centered learning environments. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent Tutoring Systems*, volume 6095 of *Lecture Notes in Computer Science* (pp. 166–177). Springer Berlin Heidelberg. 39, 149, 154, 170

[Sackett, 1987] Sackett, G. P. (1987). Analysis of sequential social interaction data: Some issues, recent developments, and a causal inference model. In J. Osofsky (Ed.), *Handbook of infant development* (pp. 878–885). New Yoerk: Wiley. 84

[Scherer, 2005] Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, 44(4), 693–727. 6, 8, 23, 35, 83, 177, 178

[Scherer & Ellgring, 2007] Scherer, K. R. & Ellgring, H. (2007). Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion*, 7(1), 113. 55, 56

[Scherer & Scherer, 1981] Scherer, K. R. & Scherer, U. (1981). Speech behavior and personality. In J. Darby (Ed.), *Speech evaluation in psychiatry* (pp. 115–135). Grune & Stratton, New York, USA. 27

[Scherer et al., 2012a] Scherer, S., Layher, G., Kane, J., Neumann, H., & Campbell, N. (2012a). An audiovisual political speech analysis incorporating eye-tracking and perception data. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (pp. 1114–1120). European Language Resources Association. 161

[Scherer et al., 2012b] Scherer, S., Marsella, S., Stratou, G., Xu, Y., Morbini, F., Egan, A., Rizzo, A., & Morency, L.-P. (2012b). Perception markup language: Towards a standardized representation of perceived nonverbal behaviors. In Y. Nakano, M. Neff, A. Paiva, & M. Walker (Eds.), *Intelligent Virtual Agents*, volume 7502 of *Lecture Notes in Computer Science* (pp. 455–463). Springer Berlin Heidelberg. 150

[Schmidt et al., 2011] Schmidt, T., Wörner, K., Hedeland, H., & Lehmberg, T. (2011). New and future developments in EXMARaLDA. In T. Schmidt & K. Wörner (Eds.), *Multilingual Resources and Multilingual Applications. Proceedings of GSCL Conference*. 65

[Schreiber et al., 2012] Schreiber, L. M., Paul, G. D., & Shibley, L. R. (2012). The development and test of the public speaking competence rubric (PSCR). *Communication Education*, 61(3), 205–233. 161

[Schutz, 1958] Schutz, W. C. (1958). *FIRO: A three-dimensional theory of interpersonal behavior.* Oxford, England: Rinehart. 24

[Shapiro, 2011] Shapiro, A. (2011). Building a character animation system. In J. Allbeck & P. Faloutsos (Eds.), *Motion in Games*, volume 7060 of *Lecture Notes in Computer Science* (pp. 98–109). Springer Berlin Heidelberg. 152

[Slater et al., 2000] Slater, M., Howell, J., Steed, A., Pertaub, D.-P., & Garau, M. (2000). Acting in virtual reality. In *Proceedings of the 3rd International Conference on Collaborative Virtual Environments* (pp. 103–110). ACM. 45

[Smith-Lovin & Brody, 1989] Smith-Lovin, L. & Brody, C. (1989). Interruptions in group discussions: The effects of gender and group composition. *American*

Sociological Review, 54, 424–435. 27

[Spence, 2003] Spence, S. H. (2003). Social skills training with children and young people: Theory, evidence and practice. *Child and Adolescent Mental Health*, 8(2), 84–96. 3, 147, 149

[Srikant & Agrawal, 1996] Srikant, R. & Agrawal, R. (1996). Mining sequential patterns: Generalizations and performance improvements. *Advances in Database Technology*, 1057, 1–17. 93

[Swartout et al., 2010] Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., Williams, J., Leuski, A., Narayanan, S., Piepol, D., Lane, C., Morie, J., Aggarwal, P., Liewer, M., Chiang, J.-Y., Gerten, J., Chu, S., & White, K. (2010). Ada and Grace: Toward realistic and engaging virtual museum guides. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Intelligent Virtual Agents*, volume 6356 of *Lecture Notes in Computer Science* (pp. 286–300). Springer Berlin Heidelberg. 1, 2

[Swartout et al., 2013] Swartout, W. R., Artstein, R., Forbell, E., Foutz, S., Lane, H. C., Lange, B., Morie, J. F., Rizzo, S., & Traum, D. R. (2013). Virtual humans for learning. *AI Magazine*, 34(4), 13–30. 38, 149

[Szekrényes, István, 2014] Szekrényes, István (2014). Annotation and interpretation of prosodic data in the HuComTech corpus for multimodal user interfaces. *Journal on Multimodal User Interfaces*, 8(2), 143–150. 69

[Talbot et al., 2012] Talbot, T. B., Sagae, K., John, B., & Rizzo, A. A. (2012). Designing useful virtual standardized patient encounters. In *Proceedings of the 2012 Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)*. 3, 39, 149

[Tamura et al., 1998] Tamura, M., Kondo, S., Masuko, T., & Kobayashi, T. (1998). Speech parameter generation algorithms for HMM-based speech synthesis. In *Proceedings of the 1998 International Conference on Acoustics, Speech, and Signal Processing (INTERSPEECH)* (pp. 3745–3748). 52

[Tan et al., 2005] Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc. 95

[ter Maat et al., 2010] ter Maat, M., Truong, K., & Heylen, D. (2010). How turn-taking strategies influence users' impressions of an agent. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Intelligent Virtual Agents*, volume 6356 of *Lecture Notes in Computer Science* (pp. 441–453). Springer Berlin Heidelberg. 145

[Thompson, 2007] Thompson, E. R. (2007). Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *Journal of Cross-Cultural Psychology*, 38(2), 227–242. 161

[Tusing & Dillard, 2000] Tusing, K. J. & Dillard, J. P. (2000). The sounds of dominance. *Human Communication Research*, 26(1), 148–171. 26

[Vaasen et al., 2012] Vaasen, F., Wauters, J., Van Broeckhoven, F., Van Overveldt, M., Daelemans, W., & Eneman, K. (2012). deLearyous: Training interpersonal communication skills using unconstrained text input. In *Proceedings of the 6th European Conference on Games Based Learning* (pp. 505–513). 78, 79

[Vilhjálmsson et al., 2007] Vilhjálmsson, H., Cantelmo, N., Cassell, J., E. Chafai, N., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A., Pelachaud, C., Ruttkay, Z., Thórisson, K., van Welbergen, H., & van der Werf, R. (2007). The behavior markup language: Recent developments and challenges. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, & D. Pelé (Eds.), *Intelligent Virtual Agents*, volume 4722 of *Lecture Notes in Computer Science* (pp. 99–111). Springer Berlin Heidelberg. 9, 104, 105, 114, 179

[Vinciarelli et al., 2009] Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12), 1743–1759. 26, 27

[Von der Puetten et al., 2010] Von der Puetten, A. M., Krämer, N. C., Gratch, J., & Kang, S.-H. (2010). "It doesn't matter what you are!": Explaining social effects of agents and avatars. *Computers in Human Behavior*, 26(6), 1641–1650. 29

[Wagner et al., 2013] Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., & André, E. (2013). The social signal interpretation (SSI) framework: multimodal signal processing and recognition in real-time. In *Proceedings of the 21st ACM International Conference on Multimedia* (pp. 831–834). ACM. 123, 125, 138

[Ward & Tsukahara, 2000] Ward, N. & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8), 1177–1207. 27, 32

[Waxer, 1977] Waxer, P. H. (1977). Nonverbal cues for anxiety: an examination of emotional leakage. *Journal of abnormal psychology*, 86(3), 306. 46

[Wiggins, 2003] Wiggins, J. S. (2003). *Paradigms of Personality Assessment*. New York: Guilford. 24

[With & Kaiser, 2011] With, S. & Kaiser, W. S. (2011). Sequential patterning of facial actions in the production and perception of emotional expressions. *Swiss Journal of Psychology*, 70(4), 241–252. 6, 32, 52, 83, 177

[Wittenburg et al., 2006] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). Elan: a professional framework for multimodality research. In *Proceedings of the 5th Language Resources and Evaluation Conference* (pp. 1556–1559). Language Resources and Evaluation. 65, 163

[Xu et al., 2014] Xu, Y., Pelachaud, C., & Marsella, S. (2014). Compound gesture generation: A model based on ideational units. In T. Bickmore, S. Marsella, & C. Sidner (Eds.), *Intelligent Virtual Agents*, volume 8637 of *Lecture Notes in Computer Science* (pp. 477–491). Springer International Publishing. 54

[Yabar & Hess, 2007] Yabar, Y. & Hess, U. (2007). Display of empathy and perception of out-group members. *New Zealand Journal of Psychology*, 36(1), 42. 8, 27, 28

Agents Conversationnels Animés pour l'entrainement social : modèle computationnel de l'expression d'attitudes sociales par des séquences de signaux non-verbaux

Mathieu CHOLLET

RESUME : Les développements dans le domaine des Agents Conversationnels Animés (ACAs), des personnages virtuels utilisant les modalités de la communication humaine pour interagir avec un utilisateur, ont permis de proposer des systèmes pour l'entrainement des compétences sociales par la pratique. De tels ACAs doivent être capables de confronter l'apprenant à toutes les situations sociales nécessaires à son apprentissage, et doivent ainsi pouvoir exprimer différentes émotions ou différentes attitudes. Les émotions et attitudes s'expriment en particulier par les signaux non-verbaux : sourires, postures, *etc*. Des recherches en psychologie ont cependant montré que l'interprétation de signaux non-verbaux est modifiée par les signaux proches : un sourire suivi d'un évitement de regard et d'un détournement de la tête n'est ainsi pas un signe d'amusement mais d'embarras. Lors de la planification des signaux non-verbaux d'un ACA, il faut donc considérer les séquences complètes de signaux non-verbaux et pas seulement ces signaux indépendamment les uns des autres. Or, la plupart des modèles existants ne prennent pas ce phénomène en compte. La contribution principale de notre thèse est de proposer une méthodologie pour extraire automatiquement d'un corpus multimodal des séquences de signaux non-verbaux caractéristiques de l'expression d'attitudes, et un modèle de planification de comportement utilisant un ensemble de séquences extraites avec cette méthode. Nous appliquons ces travaux à l'expression d'attitudes par un recruteur virtuel.

Une autre enjeu dans la conception de systèmes d'entrainement social est de vérifier si les utilisateurs améliorent bien leurs compétences sociales en s'entrainant avec de tels systèmes. Nous nous sommes intéressés à la réalisation d'une audience virtuelle constituée d'ACAs servant à s'entrainer à l'entrainement à la prise de parole en public. Notre seconde contribution est d'avoir proposé une architecture d'audience virtuelle réagissant en temps réel à la performance de l'utilisateur, et d'avoir évalué plusieurs stratégies de retour à l'utilisateur.

MOTS-CLEFS : Agents Conversationnels Animés - Entrainement Social

ABSTRACT : The Embodied Conversational Agents (ECAs) used in social training must be able to simulate all the different social situations that a learner has to train to. Depending on the application, the ECAs must then be able to express various emotions or various attitudes. Non-verbal signals, such as smiles or gestures, contribute to the expression of attitudes. However, recent findings have demonstrated that non-verbal signals are not interpretated in isolation but along with other signals : for instance, a smile followed by a gaze aversion and a head aversion does not signal amusement, but embarassment. Non-verbal behavior planning models for ECAs should thus consider complete sequences of non-verbal signals and not only signals independently of one another. However, existing models do not take this into account, or in a limited manner. The main contribution of this thesis is a methodology for the automatic extraction of sequences of non-verbal signals characteristic of attitude variations from a multimodal corpus, and a non-verbal behavior planning model that takes into account sequences of non-verbal signals rather than signals independently.

Another consideration in the design of social training systems is to check that users do improve their social skills while using such systems. We investigated the use of ECAs to build a virtual audience aimed at improving users' public speaking skills. Another contribution of this thesis is the proposal of an architecture for interactive virtual audiences that provide realtime feedback to the learner according to his public speaking performance, and to have evaluated three different feedback strategies.

KEY-WORDS : Embodied Conversational Agents - Social training