Coordination des tours de parole par le couplage sensorimoteur continu entre utilisateurs et agents

Mathieu Jégou

THÈSE / UNIVERSITÉ DE BRETAGNE OCCIDENTALE
sous le sceau de l'Université européenne de Bretagne

pour obtenir le titre de
DOCTEUR DE L'UNIVERSITÉ DE BRETAGNE OCCIDENTALE
Mention : Informatique
École Doctorale Santé, Information, Communication, Mathématique, Matière

présentée par
Mathieu JÉGOU
Préparée à l'Institut de Recherche Technologique B<>COM

Coordination des tours de parole par le couplage sensorimoteur continu entre utilisateurs et agents

Thèse soutenue le 05 octobre 2016
devant le jury composé de :

Alexandre PAUCHET
Maître de conférence HDR, Université de Normandie - LITIS - INSA Rouen / *rapporteur*

Elisabetta BEVACQUA
Maître de Conférence, ENIB-CERV/ *examinateur*

Pierre DE LOOR
Professeur, ENIB-CERV / *examinateur*

Catherine PÉLACHAUD
Directeur de Recherche CNRS, LTCI-Télécom ParisTech / *rapporteur*

Gérard BAILLY
Directeur de Recherche CNRS, GIPSA-Lab / *examinateur*

Pierre CHEVAILLIER
Professeur, ENIB-CERV / *directeur de thèse*

Remerciements

La thèse arrive à sa fin et lorsque je récapitule ces trois ans de travail je me rends compte que beaucoup de personnes ont permis à cette thèse d'aller à son terme. Je tenais tout d'abord à remercier l'IRT b-com et son directeur général Bertrand Guilbaud pour la confiance qu'ils m'ont accordée au cours des trois ans de thèse. Je mesure la chance d'avoir pu mener cette thèse dans un institut qui a toujours fait en sorte que j'aie les ressources financières et logicielles suffisantes pour arriver au bout du projet. Je tenais en particulier à remercier les membres de l'équipe des Infrastructures et Technologies. Je souhaitais remercier aussi Véronique Dupont, l'équipe de l'accueil de b-com, Karen Chapon et Marie Kollman pour m'avoir grandement simplifié mes déplacements à l'étranger, Emmanuelle Lanfray et Lucie Petta pour leur aide juridique liée à l'organisation des expérimentations, Emmanuelle Garnaud-Gamache et Delphine Jugon pour m'avoir aidé à valoriser mes publications et Marion Bénetière pour m'avoir sensibilisé à la propriété intellectuelle de mon travail.

J'ai réalisé ce projet dans une équipe pluridisciplinaire qui m'a beaucoup apporté dans les différents domaines abordés au cours de ma thèse, qui n'étaient pas forcément mes domaines de spécialité lorsque j'ai débuté cette thèse. Je pense tout particulièrement à Guillaume Jégou le directeur du laboratoire Usage et Acceptabilité, Jean-Marc Diverrez et Nicolas Jullien respectivement responsable et ancien responsable du projet UXPPE pour m'avoir aussi bien intégré au sein du projet et avoir valorisé mes travaux au sein et en dehors de b-com. Je pense aussi à tous les membres d'UXPPE, anciens ou actuels, salariés, mis à disposition ou prestataires avec qui j'ai pu travailler : Mathieu pour avoir été plus qu'un stagiaire, un ingénieur de recherche inestimable, pour avoir développé l'interface d'analyse et de visualisation de l'interaction agent-utilisateur qui est devenu une aide indispensable dans le développement de l'architecture d'agent et du modèle théorique et son aide pour la conception et la réalisation de la dernière expérimentation, Édouard pour le design de l'interface de visualisation de l'interaction utilisateur-agent, Liv pour avoir pris de son temps pour m'aider à réaliser la seconde expérimentation, Tim pour le développement de l'outil de récupération et d'agrégation de données qui m'a été indispensable dans la suite de ma thèse, Nicolas pour ses conseils avisés en statistiques, Christine pour m'avoir introduit à l'entrepreneuriat, Sébastien, Agathe, Sonia et Nico pour leurs conseils en ergonomie et sur la conception de tests uti-

lisateurs, Hugo pour m'avoir fait partager son expertise en traitement du signal, Maximilien et son expertise en infographie 3D, Karim, Virginie, Justine, Grégoire, Annabelle, Hicham, Delphine, Martin, en espérant n'avoir oublié personne.

J'ai aussi eu la chance d'acquérir énormément de connaissances dans le domaine des agents conversationnels animés grâce aux réunions du groupe « Humains Virtuels » de l'ENIB. Je souhaitais ainsi remercier Elisabetta Bevacqua et Pierre Chevaillier pour l'organisation de ce groupe et des réunions ainsi qu'à eux et aux autres membres réguliers, Pierre De Loor, Matthieu Courgeon, Alexis Nédélec, Éric Maisel, Cédric Buche, Alexandre Kabil, Mukesh Barange, Ronan Querrec avec qui j'ai beaucoup appris et de qui j'ai pu recevoir beaucoup de conseils sur la réalisation de ma thèse.

Bien sûr cette thèse n'a pu être achevée que grâce à l'implication quotidienne de Pierre Chevaillier, mon directeur de thèse. Pierre, je te remercie pour avoir été à mes côtés au cours des trois années de thèse, tu as su me permettre d'acquérir de l'autonomie et de prendre mes responsabilités de chercheur, tout en mettant les « mains dans le cambouis » quand je traversais des moments de difficultés. Je te remercie aussi pour ta bonne humeur et ta passion communicatives, et ta capacité à me motiver lors des moments de démoralisation. J'ai enfin appris énormément à tes côtés, dans le domaine des systèmes multi-agents, des architectures d'agents, et de la réalité virtuelle.

Pour terminer ces remerciements j'ai une pensée pour tous les autres salariés du site de b-com à Brest qui m'ont permis de travailler dans la bonne humeur, Hamza, Iyas, Théotime, Thomas, Hamidreza, Taman, Rafikh, César, Erwan. Je souhaite aussi, bien sûr, associer à ces remerciements, tous les volontaires de mes expérimentations sans qui ma thèse n'aurait pas pu avancer. Je voudrais, enfin, surtout, associer mes proches et mes amis, ils ont été à mes côtés, m'ont soutenu dans les moments difficiles et ont un grand rôle dans la réussite de cette thèse.

Résumé

Nous proposons dans cette thèse de résoudre la problématique de la coordination de la parole entre utilisateurs et agents conversationnels animés dans des interactions dyadiques. Selon une approche courante, coordonner la parole reviendrait à éviter des recouvrements de parole et à minimiser les moments de silence entre deux tours, ceci pour rendre plus fluide l'interaction avec l'agent et améliorer l'expérience de l'utilisateur en interaction dialogique avec l'agent. Les interactions humaines montrent néanmoins une coordination plus complexe avec des recouvrements de parole compétitifs ou non compétitifs et des moments de silences longs. Selon notre approche, c'est en permettant cette diversité des situations que nous verrons émerger, entre l'utilisateur et l'agent, une interaction plus fluide et plus crédible, améliorant l'expérience de l'utilisateur avec l'agent. Les échanges de paroles sont néanmoins, par nature, complexes, la coordination se faisant par l'interaction entre locuteur et interlocuteur plus que par un participant en particulier. Pour capturer cette complexité, nous avons élaboré un modèle mettant l'accent sur une coordination de la parole basée sur un couplage sensorimoteur continu. Nous montrons la capacité de notre modèle à faire émerger les différentes situations liées à la coordination de la parole humaine à la fois dans une interaction entre deux agents et dans une interaction utilisateur-agent.

Abstract

In this thesis, we resolve the issue of the coordination of speaking turns in dyadic interactions between users and embodied conversational agents. According to a common view, coordinating turns means avoiding overlaps and minimize silences between turns. By optimizing turn transitions between users and agents, the user's experience is expected to be improved. However, observations of human conversations show a more complex coordination of speaking turns between user and agents : awkward silences and overlaps, competitive or not, are common. In order to improve the credibility and the naturalness of the interaction, we must observe the same variability of the situations in a user-agent interaction. Nevertheless, exchanges of speaking turns are, by nature, complex, the coordination is managed by the interaction between participants more than controlled by one participant alone. To capture this complexity, we elaborated a model emphasizing the continuous sensory-motor coupling existing between the user and the agent. We show the capacity of our model to make emerge the different situations linked to the coordination of speaking turns in interactions between two agents and between one user and one agent.

Table des matières

Table des figures

Liste des tableaux

Glossaire

backchannel se réfère selon Ward et Tsukahara (2000) à des "énoncés courts" (uhm, d'accord, par exemple) ou actions (hochement de tête, par exemple) produites par l'auditeur qui répondent directement à l'énoncé du locuteur, sont optionnels, et ne requièrent pas de réponses de la part du locuteur. Ces *backchannels* peuvent être des *continuers* des énoncés ou actions incitant le locuteur à continuer ou des évaluations (*assessments*), servant à l'auditeur à renseigner son degré de compréhension. . 37

continu lorsque nous parlons de perception et d'action continues nous nous référons, d'une part, au fait que l'agent perçoit et varie ses actions à chaque instant de l'interaction (« continu » s'oppose ici à événementiel), d'autre part, au fait que l'agent perçoit ou attribue des grandeurs numériques continues aux éléments de perception (degré d'appartenance à un ensemble par exemple) et module ses commandes motrices selon une variable numérique continue (« continu » s'oppose alors à catégoriel).. 24

couplage terme employé dans différentes disciplines pour souligner l'influence mutuelle que peuvent exercer plusieurs systèmes en interaction. Lorsque nous parlons de couplages entre individus dans une interaction dialogique, nous nous référons plus précisément à la définition de De Loor *et al.* (2015) : influence mutuelle et continue entre deux personnes, qui génère une dynamique spécifique à la dyade.. 24

filler interjections (de type "euh" en français) (Clark et Fox Tree, 2002) utilisées par les participants à une conversation pour entre autres, combler un silence gênant Ohshima *et al.* (2015), ou pour l'auditeur, retarder sa prise de tour en signalant qu'il est en train de planifier la phrase qu'il va dire (Beňuš *et al.*, 2011).. 37

projection se réfère au processus par lequel un participant est capable de prédire à l'avance un moment possible de transition de tour (*Transition Relevant Place* ou *TRP*), à partir de l'identification d'unité de construction de tour tel qu'avancé par Sacks *et al.* (1974). . 58

Chapitre 1

Introduction générale

En 2040, Paul dispose, comme la majorité de la population, d'un assistant virtuel humanoïde, Isaac, possédant une représentation graphique 3D complète (corps entier) et capable de dialoguer en langue naturelle parlée avec l'utilisateur. Isaac aide Paul dans un grand nombre de décisions de la vie quotidienne.

Néanmoins, Isaac ne se contente pas de fournir des informations pratiques, mais dispose de compétences relationnelles avancées. C'est ainsi une véritable relation de confiance qui s'est créée entre Paul et Isaac. Cette relation particulière permet à Paul de se confier plus facilement à Isaac, et Isaac de mieux comprendre et s'adapter aux envies de Paul. Lorsqu'il est dans l'intérêt de Paul de réaliser une tâche, mais qu'Isaac perçoit peu de motivations de la part de Paul, il exploite sa connaissance avancée de la personnalité de Paul pour le convaincre. Ceci est d'autant plus important que Paul souffre de problèmes de santé l'obligeant à prendre un traitement médicamenteux. Étant donné la nature du traitement, Paul est parfois démotivé à le prendre et c'est Isaac qui persuade Paul de le poursuivre son traitement. Imaginons justement une situation où Paul, en retard à son travail décide de ne pas prendre son traitement. Isaac sait qu'il est important que Paul prenne son traitement à heure régulière, et s'inquiète du fait que la décision de Paul puisse entrainer des prises de plus en plus irrégulières de son traitement. Juste avant de partir, Isaac lui rappelle qu'il n'a pas pris son traitement. Le début de l'échange est transcrit dans l'encart 1.1.

S'ensuit une série d'arguments de la part d'Isaac sur les risques pour la santé de Paul et sur la probabilité que la décision de Paul ce jour là entraîne d'autres décisions similaires à l'avenir.

Par ce scénario, nous avons imaginé ce que pourrait être une interaction dyadique naturelle et entre un utilisateur et un agent conversationnel animé. Les agents conversationnels animés fascinent de nombreux chercheurs depuis maintenant plus de 20 ans et ceux-ci n'ont cessé d'améliorer les capacités dialogiques, affectives et émotionnelles de ces personnages humanoïdes parlants. D'abord cantonnés aux laboratoires de recherche, les premiers systèmes apparaissent maintenant sur le marché et sont promis à un avenir radieux (Jump et Ekholm, 2015) comme assistants per-

```
(1) Isaac : Paul je vois que tu es en train de partir, as-tu pris ton
     traitement ?
(2) Paul : (0.5) Eu:h non, je suis en retard au travail, tant [pis je le
     prendrai demain.]
(3) Isaac :                                                    [Tu sais
     qu'il faut que tu sois régulier] dans la prise de ton traitement.
(4) Paul : (0.4) un jour sans le prendre ce n'est pas dramatique.
(5) Isaac : (0.5) Il faut que tu le prennes régulièrement, [Paul, sinon
     ...sinon ]
(6) Paul :                                                 [Je suis
     capable]
(7) Isaac : ta santé eu:h [risque de se ] dégrader.
(8) Paul :               [je peux ...]
(9) Isaac : (3.0) [je ...]
(10) Paul :       [et ...]
(11) Isaac :  (1.0) enfin je cherche juste à m'assurer que tu le prennes
     bien, et que si tu ne le prends pas tu as conscience des risques
     que tu encours.
```

TABLE 1.1 – Extrait d'une interaction entre Paul et Isaac.

sonnels, coéquipiers, tuteurs artificiels ou encore comme agents d'accueil de sites web.

Dans le scénario présenté dans l'encart 1.1, nous avons plus particulièrement souhaité mettre en avant les moments de prise de parole de Paul et Isaac. Pour cela nous avons renseigné les temps de prise de parole et les interruptions selon les notations introduites par Jefferson (2004). Dans ce système de notation, un nombre décimal entre parenthèses au début du tour d'un participant indique la durée de silence avant la prise de parole, tandis qu'un nombre entre parenthèses pendant le tour d'un participant indique une pause. Le symbole « [» indique le début d'un recouvrement de parole entre deux participants, «] » indique la fin du recouvrement. Le symbole « : » indique enfin la prolongation de la prononciation du phonème précédent.

Nous montrons, par cette interaction, le caractère fluide et naturel des échanges de parole. Dans ce scénario, nous présentons, dans l'ordre, une prise de parole suivant une période de silence de 500 ms après la fin de tour de l'agent (ligne (2) du tableau 1.1), une interruption de l'agent (ligne (3)), une prise de parole suivant un court moment de silence de l'agent (ligne (5)), plusieurs tentatives d'interruption de l'utilisateur (lignes (6) et (8)) et un long moment de silence suivi d'une prise de parole simultanée et involontaire des deux participants (lignes (9) et (10)). Dans toutes ces situations, les participants se coordonnent finement pour résoudre les conflits de parole (moments où les deux participants cherchent à avoir la parole en même temps) ou des moments de silence trop longs. Une telle forme d'interaction est un enjeu dans le développement des agents conversationnels animés. En effet, les interactions courantes apparaissent encore trop rigides, saccadées, l'agent ne pouvant interpréter la phrase de l'utilisateur que lorsque ce dernier a fini de la prononcer (Kopp *et al.*, 2014). Lorsque l'agent dispose de capacités à interpréter et

réagir à l'énoncé de l'utilisateur avant que ce dernier ait fini, ce sont les concepteurs de l'agent qui imposent le tour par tour strict, en cherchant à doter l'agent de capacités à optimiser la détection de la fin de tour de l'utilisateur de sorte de prendre le tour le plus rapidement possible en évitant tout recouvrement de parole avec l'utilisateur (Jonsdottir et Thórisson, 2013). Cette vision d'un échange de parole stricte et sans coupures, composé de règles formelles à suivre par les participants se base sur le modèle de Sacks *et al.* (1974). Néanmoins, des auteurs plus récents tels que O'Connell *et al.* (1990) et Clark (1996) ont critiqué ce modèle en montrant que les comportements liés au tour de parole sont plus variés que proposé initialement par Sacks *et al.* (1974). En effet, dans des conversations spontanées humaines, il n'est pas rare d'observer des interruptions et des moments de silence longs, ces événements pouvant être liés à la nature de la contribution verbale du participant (Clark, 1996) ou des attitudes interpersonnelles de ces derniers (Ter Maat *et al.*, 2010; Ravenet *et al.*, 2015).

Si les interruptions sont communes dans les conversations humaines, nous pouvons nous attendre, dans une interaction agent-utilisateur, à observer les mêmes comportements de la part de l'utilisateur. En cas de tentatives d'interruptions ou de prises de paroles simultanées, selon la nature de l'information qu'il transmet, il garde ou laisse le tour à l'utilisateur. Dans le scénario, lorsque l'utilisateur cherche à interrompre l'agent en prononçant le début de phrase « Je peux », ligne (8), l'agent continue à parler malgré la tentative d'interruption de l'utilisateur, jugeant l'information qu'il souhaite transmettre trop importante pour l'interrompre. Cette tentative d'interruption avortée est suivie d'un moment de réflexion de l'agent qui modifie le cours de la prononciation de sa phrase tel que le montre le terme « eu :h », ligne (7). De même le choix par l'agent d'interrompre l'utilisateur lorsqu'il prononce l'énoncé « Tu sais qu'il faut que tu sois régulier dans la prise de ton traitement » ne devrait pas être conçu comme une erreur, ou une violation des règles d'interaction avec l'utilisateur, mais provenant d'une intention communicative particulière de l'agent.

Si les interruptions volontaires ne sont pas rares, les paroles simultanées involontaires ont inévitablement lieu dans les interactions humaines (Sacks *et al.*, 1974) dues à des situations où les participants n'ont pas de contraintes sur le moment où ils peuvent prendre la parole. Nous montrons un exemple de de paroles simultanées involontaires aux lignes (9) et (10). Après les tentatives d'interruption de l'utilisateur au tour précédent, chaque participant s'attend à ce que l'autre prenne le tour. Il en résulte alors une situation d'attente résultant en un moment de silence long (3 secondes), puis les deux participants, constatant que leur partenaire respectif ne prend pas le tour, cherchent à relancer la conversation en prenant le tour. Percevant alors la parole simultanée, l'agent et l'utilisateur s'interrompent pour laisser la parole à l'autre et de nouveau un silence apparait qui est cette fois suivie par une prise

de parole de l'agent.

Nous pouvons nous interroger sur l'intérêt de doter un agent de capacités à coordonner ses tours de parole de manière similaire à ce que nous pouvons observer dans les interactions humaines.

D'une part, notre objectif ne réside pas dans une optimalité de la perception du comportement de l'utilisateur mais plutôt dans la crédibilité de l'agent dans l'interaction. Nous souhaitons un agent doté des mêmes capacités de perception et d'action qu'un humain de sorte que si une situation amène un participant humain à interrompre involontairement le locuteur courant, le même comportement doit être observé chez l'agent dans une situation similaire.

D'autre part, toutes les situations liées à la coordination de la parole dépendent de la contribution des deux participants. Lors d'une fin de tour, le locuteur courant signale activement sa fin de tour à l'auditeur, ce dernier identifiant ces signaux pour prendre le tour. La prise de tour n'est donc pas le seul fait de l'auditeur : si le locuteur courant ne fournit pas de signaux de fins de tour, l'auditeur sera incertain sur le fait que le locuteur courant a fini son tour ou effectue simplement une pause dans son discours. De même, en situation de conflit, les participants échangent des signaux révélant leur intention de garder ou laisser la parole. Ces signaux aident les participants à résoudre de manière fluide les situations de conflit. Si l'un des deux participants ne fournit pas de tels signaux, son partenaire sera incertain sur le fait qu'il s'apprête à laisser ou garder le tour à un instant donné allongeant grandement la durée du conflit. Il est donc nécessaire pour un agent conversationnel, non seulement d'interpréter correctement les signaux de l'utilisateur mais aussi de se rendre compréhensible de son partenaire et la manière la plus naturelle de le faire est de reproduire les signaux verbaux et non verbaux existants dans les interactions humaines.

La nécessité, non seulement d'interpréter correctement les actions du participant, mais aussi de se faire comprendre est un indicateur du couplage sensorimoteur continu existant entre les partenaires. En effet, les participants sont en situation de co-dépendance dans l'évolution de la production de leurs signaux. Qu'à un instant donné, en situation de conflit, un participant produise des signaux laissant penser qu'il s'apprête à laisser la parole ou non impactera la production de signaux de l'autre participant : ce dernier accentuera ou non ses propres signaux indiquant qu'il garde la parole. Aussi la production de signal de chaque participant ne peut être prédite à l'avance, elle émerge de l'interaction entre les partenaires. L'idée d'un couplage continu entre les participants est en rupture avec les modèles de coordination du tour de parole existants mais a été étudiée pour d'autres formes d'interactions par Ikegami et Iizuka (2007) et Bevacqua *et al.* (2014). Pour la réalisation de notre modèle nous nous inscrivons dans un paradigme de perception-action similaire aux approches prises par ces auteurs et en s'inspirant des travaux de psychologie cogni-

tive de Marsh *et al.* (2006), Warren (2006) et Kelso (2009).

La nature de notre modèle nous laisse avec quatre questions de recherche que nous tachons de résoudre dans cette thèse.

Nous cherchons tout d'abord à évaluer l'intérêt d'un modèle continu et émergent capable de reproduire la coordination des tours de parole observée dans les interactions humaines. Aussi, notre intérêt n'est pas d'accroître la performance de l'agent à réagir de manière optimale aux fins de tour de l'utilisateur mais plutôt de doter l'agent de la capacité à maintenir la coordination des échanges de parole avec l'utilisateur. Si l'utilisateur produit un comportement que l'agent ne connait pas, ce dernier est capable de s'adapter rapidement pour revenir à une situation stable dans les échanges de parole. Nous souhaitons donc répondre à la question suivante au cours de cette thèse :

Question générale 1. *Une approche émergente à base d'un modèle continu permet-elle de reproduire les comportements relatifs à la gestion des tours de parole dans des interactions humaines ?*

Les interactions dialogiques sont de même dynamiques : le comportement de chaque participant et l'état de l'environnement varient continuellement au cours de la conversation. Dans les interactions humaines, les partenaires sont alors capables de s'adapter rapidement à ces variations. De plus, l'environnement est souvent bruité, dégradant la capacité à percevoir les variations de comportement du partenaire. Malgré tout, les participants humains sont capables de s'adapter au caractère bruité de l'environnement, et de garder une coordination effective de la parole dans ces situations.

Nous souhaitons ainsi vérifier que notre agent possède les mêmes capacités d'adaptations aux variations du comportement de son partenaire et à différentes conditions environnementales. Nous introduisons donc la question suivante :

Question générale 2. *À quel point notre agent est capable de s'adapter à son partenaire et à un environnement bruité ?*

Nous avons postulé l'importance d'un agent ne se contentant pas de prendre ou laisser le tour à l'utilisateur mais reproduisant le plus exhaustivement possible les comportements observés dans les interactions humaines. Nous pensons ainsi que de tels comportements pourrait modifier la perception que l'utilisateur a d'un agent disposant d'un répertoire de comportements variés par rapport à un agent contrôlé par un modèle plus traditionnel de coordination de la parole. Nous nous posons donc la question suivante :

Question générale 3. *Un agent capable de reproduire la variabilité des situations liées au tour de parole modifie-t-il la perception que l'utilisateur a de l'agent ?*

Enfin l'application d'un paradigme continu pour la gestion de l'interaction utilisateur-agent est un défi à part entière : notre approche est en effet en rupture avec les approches événementielles existantes. Pour cette raison, les architectures d'agents conversationnels actuelles ne sont pas adaptées à des modèles continus. Pour implémenter notre modèle, nous devrons donc au minimum adapter une architecture existante et démontrer la capacité de notre implémentation à fonctionner dans le cadre d'une interaction temps-réel avec l'utilisateur. Nous nous posons donc la question suivante :

Question générale 4. *À quel point notre modèle de gestion du tour de parole peut être adapté à un contexte réel de dialogue utilisateur-agent ?*

Notre contribution consiste en la conception du modèle conceptuel pour la gestion du tour de parole (chapitres 6 et 7), la création d'une architecture d'agent pour l'application de ce modèle à une interaction temps réel utilisateur-agent (chapitres 8 et 10) et en sa comparaison avec des interactions humaines (chapitre 9) et des interactions temps réel entre un agent piloté par notre modèle et l'utilisateur (chapitre 10).

Première partie

État de l'art

Chapitre 2

Agents conversationnels animés

2.1 Définition

Les agents conversationnels animés sont des entités informatiques capables de communiquer avec l'utilisateur en utilisant le langage naturel et la communication non verbale (André *et al.*, 2005). Ils apparaissent à l'utilisateur comme des personnages 2D ou 3D humanoïdes disposant d'une représentation permettant l'emploi des mêmes signaux non-verbaux qu'un humain (Berry *et al.*, 2005). Ils disposent enfin d'un degré d'autonomie leur permettant de prendre des décisions de manière proactive sur les actions à réaliser (Berry *et al.*, 2005). Le développement des agents conversationnels animés a connu un intérêt grandissant depuis les années 1990 (Elliott et Brzezinski, 1998). D'une part, par un dialogue en langue naturelle parlée, l'utilisateur se passe des interfaces traditionnelles nécessitant un apprentissage de sa part (Cassell *et al.*, 1999). Ensuite, la communication utilisateur-agent est potentiellement plus robuste. En effet le caractère multimodal de l'interaction permet une redondance dans la transmission d'informations : si l'on ne réussit pas à interpréter les signaux transmis par l'utilisateur sur un canal de communication particulier, on peut se fier à d'autres canaux de communication pour interpréter les intentions communicatives de l'utilisateur (Cassell *et al.*, 1999). Enfin, cet intérêt coïncide, d'une part avec l'apparition de l'informatique affective (Picard, 2000), domaine dédié au traitement et la génération d'émotions par un programme informatique, et au développement du paradigme « Computers As Social Actors » (CASA) de Reeves et Nass (1996). À l'origine de ce paradigme, Reeves et Nass (1996) proposent une série d'expérimentations montrant qu'un utilisateur tend à percevoir un média comme une entité sociale. L'utilisateur est ainsi capable d'attribuer à des entités artificielles un caractère plus ou moins extraverti, plus ou moins poli ou plus ou moins amical. L'emploi d'agents conversationnels animés présente un grand intérêt vis-à-vis de ce paradigme : en observant et modélisant la manière dont les humains interagissent on peut susciter chez l'utilisateur des actions ou réactions particulières permettant de favoriser le caractère naturel de l'interaction, l'immersion de l'utilisateur ou encore

son engagement avec le système (Elliott et Brzezinski, 1998).

2.2 Gestion des conversations dans les interactions humaines

Les agents conversationnels animés interagissent avec l'utilisateur par un mode d'interaction reproduisant les conversations humaines. Afin de comprendre les problématiques liées à la conception de ces interfaces, il est donc nécessaire de comprendre la manière dont se déroulent les interactions humaines. C'est l'objet de cette section. Goffman (1976) propose deux définitions différentes d'une conversation :

1. « des échanges verbaux informels dans des interactions quotidiennes » différant d'autres formes d'interactions verbales plus formelles comme une présentation orale ou un débat politique dans lesquels les temps de parole des participants sont contrôlés par une personne tiers ;
2. toute forme d'interaction verbale.

De manière similaire à Sacks *et al.* (1974) et Clark (1996) lorsque nous employons dans ce mémoire le terme de conversation nous nous référons à la première définition. Néanmoins contrairement à Goodwin (1981) et Clark (1996) considérant tous deux que la conversation est nécessairement face à face, nous considérons comme Sacks *et al.* (1974) une variété d'interactions médiatisées ou non, allant des conversations téléphoniques aux interactions face à face.

La principale caractéristique d'une conversation provient du fait qu'elle est d'initiative mixte : aucun participant en particulier n'a seul la décision sur le déroulé de l'interaction, mais celle-ci est distribuée entre les participants, et chaque participant peut librement et à tout moment apporter sa contribution verbale à l'interaction (Clark, 1996). Ainsi sont exclues du champ des conversations des interactions telles que des entretiens où un participant a un ensemble de questions qu'il pose dans un ordre donné à un ou des interlocuteurs indépendamment des réponses que ces derniers produisent (Clark, 1996).

La conversation n'est ainsi pas organisée à l'avance mais émergente, les actions sont co-construites localement par l'interaction entre les participants. La conversation est alors le résultat de l'ensemble des actions locales produites par les participants (Clark, 1996).

De plus la conversation n'est pas le simple échange d'informations verbales entre les participants. Plusieurs formes de coordination sont en effet observées.

1. Les participants s'assurent qu'il y a une compréhension mutuelle du contenu de la conversation (processus de *grounding*). Les interlocuteurs produisent par exemple des actions montrant qu'ils sont engagés dans l'interaction, qu'ils

sont capables de percevoir ce que dit le locuteur et qu'ils sont capables de comprendre ce qui est dit (Clark, 1996).

2. Les participants assurent une coordination des échanges de parole de sorte qu'un seul participant parle à la fois (Sacks *et al.*, 1974).
3. Des coordinations posturales (Fowler *et al.*, 2008) et des alignements de la hauteur de voix, du niveau de langage, de la vitesse de parole (Giles et Coupland, 1991), souvent regroupées sous le terme générique « synchronie » (Delaherche *et al.*, 2012) sont observés.

Pour se coordonner à ces différents niveaux, les participants s'échangent des informations transmises sous la forme d'actions appelées des signaux (Clark, 1996). Les participants transmettent ces signaux par différents canaux de communication (Goodwin, 1981). On distingue ainsi les informations verbales, information transmise par l'emploi de mots, des informations appartenant au langage corporel, signaux produits par des canaux de communication autres que la voix et le langage para-verbal, représentant la variation des attributs de la voix comme la hauteur de voix, le volume sonore ou encore la vitesse de parole. Par souci de simplicité, nous regrouperons le langage corporel et le langage para-verbal dans la catégorie du langage non-verbal en accord avec la définition du langage non-verbal proposée par Hecht et Ambady (1999).

Selon le dispositif conversationnel, les processus en jeu dans la coordination des participants ne sont pas tout à fait les mêmes. Ainsi on peut distinguer les interactions dyadiques (entre deux participants) et les interactions multipartites (trois participants ou plus). Dans les interactions dyadiques, deux rôles sont définis, le locuteur et l'auditeur (Clark, 1996). Dans les interactions multipartites, le locuteur peut à certains moments de la conversation choisir de s'adresser seulement à une partie des participants de la conversation qui deviennent les interlocuteurs, les autres participants étant, soit des *side participants* (participant qui n'est pas le destinataire mais qui doit suivre l'énoncé du locuteur) soit des *overhearers* écoutant la conversation mais n'étant pas reconnus comme participants à la conversation (Clark, 1996). Dans le cadre de ce mémoire, nous considèrerons uniquement le cas de conversations dyadiques.

2.3 Conception d'agents conversationnels animés

La conception d'architectures d'agents conversationnels animés et leur évaluation pose un certain nombre d'enjeux que nous détaillons ci-dessous. Nous proposons tout d'abord une revue de l'état de l'art de diverses dimensions d'expérience utilisateur habituellement attribuées aux agents conversationnels animés. Nous nous intéressons ensuite aux problématiques d'implémentation d'un agent conversationnel animé pour la gestion de la multimodalité et de l'interaction avec l'utilisateur

en temps-réel.

2.3.1 Expérience utilisateur et agents conversationnels animés

De nombreux auteurs ont montré l'avantage de l'emploi d'agents conversationnels animés dans plusieurs contextes d'interaction (Bickmore *et al.*, 2011; Lucas *et al.*, 2014). Employés en tant qu'agents pédagogiques, ils améliorent la mémorisation et l'apprentissage (Bickmore *et al.*, 2011), offrent une confiance accrue dans le système (Lucas *et al.*, 2014) ou encore aident à la promotion d'exercices physiques (Bickmore *et al.*, 2010).

Engagement

Une condition nécessaire pour permettre une interaction efficace et réussie entre l'utilisateur et l'agent est la capacité de ce dernier à susciter l'engagement de l'utilisateur (Peters *et al.*, 2005). La notion d'engagement comporte néanmoins plusieurs dimensions et définitions différentes selon les auteurs (Chapman *et al.*, 1999; Brockmyer *et al.*, 2009; Boyle *et al.*, 2012). Nous exposons ici les principaux points d'accord entre auteurs sur les caractéristiques de l'engagement et la manière de susciter l'engagement d'un utilisateur. Afin de donner une définition simple et consensuelle de l'engagement, nous pouvons considérer qu'un utilisateur engagé a son attention focalisée sur l'interaction, éprouve du plaisir et un intérêt à interagir avec l'interface (O'Brien et Toms, 2008). L'engagement a été étudié dans de nombreux types d'interface et a montré être une des causes du succès ou de l'échec d'un système informatique (Boyle *et al.*, 2012) et un facteur d'amélioration de l'apprentissage dans le domaine des jeux sérieux (Chapman *et al.*, 1999).

Une notion connexe existe dans le cadre des interactions dialogiques humaines : l'engagement conversationnel. Cette forme d'engagement a deux dimensions. Une première, est liée à la motivation des participants à interagir avec le ou les autres participants. Suivant cette dimension, Peters *et al.* (2005) définissent l'engagement dans le cadre d'une interaction dialogique humaine ou avec un agent conversationnel comme « la valeur qu'un participant attribue à être avec l'autre participant et à continuer l'interaction ». Une seconde dimension est liée à la manière dont les participants montrent qu'ils sont toujours présents et attentifs à ce que dit l'agent. Sidner *et al.* (2004) définissent ainsi l'engagement conversationnel comme « le processus par lequel deux participants ou plus établissent, maintiennent et finissent leur interaction ».

Peu importe le type d'engagement étudié (avec un média ou spécifiquement avec un agent conversationnel animé), plusieurs caractéristiques clés se retrouvent dans les différentes définitions. L'engagement pourrait ainsi se résumer à un sentiment de

« connexion » avec une autre entité (Sidner *et al.*, 2004), ce sentiment de connexion s'exprimant par l'attention et l'implication dont les participants font preuve. Dans la majorité des définitions, l'engagement comporte trois phases (Sidner *et al.*, 2004; O'Brien et Toms, 2008) : une phase d'initiation comprenant les motivations initiales à s'engager et les actions entreprises pour démarrer l'interaction, une phase de maintien où des actions sont produites dans le but de signifier son engagement ou garder l'autre engagé et une phase de terminaison comprenant des actions servant à conclure l'interaction.

L'engagement est un sujet particulièrement étudié dans le cadre des interactions utilisateur-agent. L'engagement est favorisé par l'emploi de signaux liés au processus d'engagement conversationnel (Rich *et al.*, 2010). Ils comprennent des regards faciaux mutuels (regards des participants l'un vers l'autre), ou des regards partagés (regard vers un objet pointé par l'autre participant) ou encore la production de *feedbacks* liés au processus de *grounding* (Rich *et al.*, 2010). L'emploi de ces signaux est nécessaire mais pas suffisant pour permettre l'engagement de l'utilisateur. Dans le cadre d'agents relationnels, des agents accompagnant l'utilisateur sur une période de temps de plusieurs mois, ces signaux ne suffisent pas et d'autres éléments sont à prendre en compte. L'agent peut aussi employer des stratégies d'établissement de rapports avec l'utilisateur (Bickmore et Picard, 2005). Cela permet de profiter de nombreuses recherches réalisées en psychologie sociale sur la manière dont deux personnes établissent un rapport (amical par exemple) à long terme (Bickmore et Picard, 2005). Bickmore et Picard (2005) proposent l'implémentation d'un agent destiné à interagir tous les jours pendant dix minutes avec un utilisateur dans l'objectif de promouvoir une activité physique régulière. Pour établir une relation à long terme avec l'utilisateur, l'agent utilise des dialogues sociaux (appelés *smalltalk*) et finissent une interaction en faisant référence à la prochaine interaction (demander quand l'utilisateur reprendra contact avec l'agent, par exemple). En plus de ces dialogues, l'agent propose des expressions d'empathie et d'humour dans le but de faciliter l'établissement d'un rapport à long terme avec l'utilisateur.

L'engagement a aussi été étudié de manière plus générale dans le cadre d'interactions entre l'utilisateur et un média. Dans ces approches, il est ainsi particulièrement mentionné qu'un utilisateur engagé sent un équilibre entre un sentiment de compétence vis-à-vis de la tâche à accomplir et sa capacité à percevoir la manière dont il peut améliorer ses compétences vis-à-vis de l'interface (O'Brien et Toms, 2008; Rozendaal *et al.*, 2009).

Rozendaal *et al.* (2009) modélisent l'engagement de l'utilisateur comme un équilibre entre la richesse du média et la sensation de contrôle perçue par l'utilisateur. La richesse du média est définie comme la variété de pensées, d'actions et de perceptions offertes par un média. Ainsi, plus une interface proposera de modalités sensorielles, de fonctions et de possibilités d'action plus la richesse sera grande. L'autre facteur

de l'engagement est le niveau de contrôle perçu par l'utilisateur. Le contrôle perçu est défini comme la perception de l'utilisateur sur sa capacité à accomplir ses buts par l'intermédiaire du média. Le contrôle est lié à des facteurs comme la facilité ou la clarté, l'efficacité ou la confiance en soi, le développement des compétences et la liberté d'action.

Selon le même principe d'équilibre entre sentiment de compétence et richesse de l'interface, certains jeux-vidéos adaptent leur niveau de difficulté au comportement de l'utilisateur afin de maintenir l'équilibre entre niveau de difficulté et compétences de l'utilisateur (Aponte *et al.*, 2011).

Crédibilité

Un des facteurs les plus traités dans l'évaluation d'interactions utilisateur-agent est la crédibilité. La crédibilité est communément définie comme « la suspension consentie de l'incrédulité » envers le fait que l'utilisateur interagit avec une entité informatique et non-sociale (Hayes-Roth et Doyle, 1998), et « l'illusion de vie » créée par le personnage virtuel (Ortony, 2003). La crédibilité semble un objectif en soi pour certains auteurs, liée dans certains cas à la capacité d'un agent à tromper l'utilisateur en lui faisant croire qu'il est humain. La crédibilité est aussi liée à la capacité d'identifier l'intention de l'agent à réaliser un acte communicatif. Il est ainsi plus difficile d'identifier les signaux verbaux et non-verbaux produits par un agent peu crédible (Pinchbeck, 2008), et de ce fait s'engager dans le dialogue. Au delà de cette définition générale, la crédibilité est une notion multidimensionnelle prenant en compte la compétence perçue de l'agent (Burgoon *et al.*, 2000), la capacité des utilisateurs à faire correspondre une action à des croyances, désirs ou intentions de l'utilisateur (Riedl et Young, 2005), ou encore le sentiment de « chaleur » (comprenant le caractère amical, la sincérité et la confiance) induit par l'agent (Niewiadomski *et al.*, 2010).

Ces dimensions montrent ainsi le potentiel d'un agent crédible à améliorer le sentiment d'aise dans l'interaction, ou encore le pouvoir de persuasion de l'agent envers l'utilisateur (Lee et Nass, 2003). Les agents crédibles sont considérés comme nécessaires dans le domaine de la narration interactive, où la compréhension de l'histoire réside dans la capacité de l'utilisateur à croire l'histoire (Riedl et Young, 2005).

La crédibilité étant de première importance pour le succès de l'interaction entre un agent conversationnel animé et un utilisateur, plusieurs études ont porté sur les facteurs permettant à un agent d'être plus crédible. L'agent virtuel doit posséder, en plus de capacités de perception et d'action humaines, un comportement cohérent entre la parole, les comportements non-verbaux et l'apparence en plus de produire des actions de manière cohérente avec la situation et les attentes de l'utilisateur (Niewiadomski *et al.*, 2010).

Le réalisme est un concept proche de la crédibilité et se réfère, dans le cadre des agents conversationnels animés à la capacité de ces derniers à reproduire l'humain par leur représentation graphique et leur comportement. Le réalisme est ainsi fortement lié à la qualité de la représentation graphique de l'agent (van Vugt *et al.*, 2007), ainsi qu'à la capacité de l'agent à reproduire dans une situation des comportements observés dans les interactions humaines (Groom *et al.*, 2009). Contrairement à un agent réaliste, un agent crédible n'a pas obligation d'être une reproduction d'êtres humains, ni en termes de représentation graphique ni en termes de comportement (Riedl et Young, 2005). Il suffit pour un agent crédible que ses comportements soient perçus comme liés à une intentionnalité et apparaissent cohérents par rapport aux intentions communicatives de l'agent et aux attentes de l'utilisateur.

Présence sociale

Beaucoup d'interactions sont aussi évaluées sur le degré de présence sociale. La présence sociale comporte plusieurs définitions à la fois liées aux technologies employées dans le cadre d'interactions humaines médiatisées et aux comportements et perceptions des participants à une interaction sociale médiatisée ou non. Lombard et Ditton (1997) considèrent la présence sociale comme une dimension de la présence. La présence est définie par ces derniers comme une illusion de non-médiation : un utilisateur agit et réagit aux événements d'un environnement virtuel comme s'il se situait dans un environnement réel. La présence sociale, elle, est définie comme le sentiment « d'être ensemble » dans un environnement partagé (Lombard et Ditton, 1997) ou encore d'être avec une autre entité sociale (Biocca *et al.*, 2003). Tel que rapporté par Biocca *et al.* (2003), certains auteurs considèrent la présence physique comme une variable binaire : un participant est présent ou non dans un environnement. D'autres auteurs considèrent la présence sociale comme une variable continue. Dans cette vision, la présence sociale est liée à la disponibilité des canaux de communication de chaque participant. Selon l'accessibilité des canaux de communication et la capacité d'un médium à reproduire fidèlement les signaux verbaux et non verbaux des participants le sentiment de présence sociale sera plus ou moins fort (Short *et al.*, 1976). D'autres auteurs raffinent la définition de présence sociale en y ajoutant la notion de conscience mutuelle : les participants ont conscience de la capacité de leur interlocuteur à percevoir leur présence et à y réagir (De Greef et Ijsselsteijn, 2000). La présence sociale n'est alors plus seulement qu'un indicateur du caractère perceptible des signaux de l'autre, mais relève de la capacité d'un participant à « accéder à l'intelligence de l'autre » (Biocca, 1997) : à percevoir dans les actions de l'autre un caractère social et une intentionnalité. À cette définition sont souvent associées des caractéristiques liées à l'engagement du participant et au sentiment de similarité dans les attitudes, émotions et comportements des participants. Bailenson et Yee (2005) montrent en ce sens qu'un agent mimant les comportements de l'utilisa-

teur génère un sentiment de présence sociale plus grand qu'un agent ne mimant pas l'utilisateur. Cet effet est appelé effet caméléon.

Dans le cadre d'interactions utilisateur-agent, Lee et Nass (2003) montrent qu'un sentiment de présence sociale accrue a des effets positifs sur la capacité de persuasion d'agents employés sur des sites de commerce en ligne.

2.3.2 Problématiques de conception d'agents conversationnels animés

Interaction multimodale temps-réel

Prenons l'exemple d'un agent et d'un utilisateur engagés dans une négociation animée pour déterminer le prix d'un article vendu par l'agent. L'agent énonce son argument justifiant à l'utilisateur le prix de l'article qu'il souhaite lui vendre. L'utilisateur mécontent n'attend pas que l'agent ait fini sa phrase et l'interrompt. L'agent détecte l'interruption de l'utilisateur. En réponse, il interrompt le cours de sa phrase, répète la dernière syllabe qu'il prononçait avant la tentative d'interruption pour signifier à l'utilisateur qu'il souhaite garder le tour, mais devant l'insistance de l'utilisateur il choisit de s'interrompre.

Ce type de scénario est courant dans le cadre d'interactions humaines et provient d'un conflit entre différents processus de coordination (Thórisson, 2002). L'agent est engagé dans un processus argumentatif où il cherche à convaincre l'utilisateur de son choix, quand l'utilisateur cherche à l'interrompre. Cette interruption est détectée par les composantes de l'agent responsables des échanges de tour qui interrompent en retour la prononciation de la phrase de l'agent pour gérer l'interruption avec l'utilisateur. Le tout se fait en quelques centaines de millisecondes avant que l'agent ne s'interrompt pour laisser la parole à l'utilisateur. Ce scénario particulier montre un exemple des contraintes auxquelles est soumis un agent pour assurer un dialogue fluide et efficace. La capacité à interrompre une action en cours pour réaliser une autre prioritaire est essentielle pour un agent destiné à avoir une interaction temps réel et manquer de réagir de manière appropriée peut conduire à l'arrêt prématuré de l'interaction. Dans l'exemple ci-dessus, si l'agent n'avait pas réagi à temps, l'utilisateur aurait déduit que l'agent ne tenait pas compte de son interruption, diminuant le caractère naturel et crédible de l'interaction, et le sentiment que l'agent prend en compte ses comportements. Le caractère temps réel de l'interaction utilisateur-agent nécessite que l'agent fasse un compromis entre rapidité de réponse et caractère approprié de sa réponse (Thórisson, 2002). Il ne prend ainsi pas toujours de décisions optimales sur l'action à réaliser mais sa réaction dépend de l'action à réaliser et des contraintes qui lui sont soumises. Plus généralement, un participant humain, ou artificiel, est soumis à plusieurs contraintes de coordination s'effectuant à différentes échelles de temps. Pour un participant humain, le temps que nécessite l'interpréta-

tion d'un stimuli visuel complexe (une image par exemple) et sa description verbale est de l'ordre de 1500 ms (Torreira *et al.*, 2015). Or, d'autres processus de coordination s'effectuent à des échelles de temps moindres, les conflits de parole dans une conversation sont par exemple très souvent résolus en moins d'une seconde (Sacks *et al.*, 1974). Cela nécessite donc l'utilisation de processus de coordination agissant en parallèle et de manière concurrente avec différents cycles de perception-action permettant, comme dans l'exemple donné au début de cette section, une réaction rapide à des événements imprévus par l'agent (Thórisson, 2002).

Multimodalité

Imaginons un agent cherchant à indiquer à l'utilisateur un itinéraire pour aller d'un point A à un point B. L'agent lui explique comment se rendre au point B en désignant du doigt les différents endroits sur la carte par lesquels il devra passer (1). Il vérifie que l'utilisateur est toujours attentif à ce qu'il dit en détectant les *backchannels* que l'utilisateur produit (2), c'est à dire des énoncés courts ou des actions renseignant sur le degré de compréhension ou d'accord de l'auditeur (Ward et Tsukahara, 2000). S'il voit que l'utilisateur ne suit pas ce qu'il dit, il réfléchit à une autre manière d'expliquer l'itinéraire. Ne sachant pas tout de suite comment mieux formuler son explication, il effectue une pause dans son discours : il signale cette pause à l'utilisateur en détournant son regard et en produisant un filler (3) tel que la production d'un « euh ».

Dans les conversations, la transmission d'information se fait de manière multimodale : les informations transmises par les participants se font sur plusieurs canaux de communication. L'agent présenté dans le scénario ci-dessus intègre l'environnement partagé par l'utilisateur en pointant du doigt les différents éléments auxquels il fait référence dans son énoncé. De même, dans le cas (3), en choisissant de produire conjointement un *filler* et en détournant son regard, l'agent s'assure que l'utilisateur n'interprète pas la pause dans son énoncé comme une opportunité à prendre la parole. De plus, une même information peut être transmise au choix par plusieurs canaux de communication permettant de s'adapter à différents types d'environnement et aux participants (Bevacqua *et al.*, 2010). Si la disposition de l'environnement est inadaptée à la transmission par une modalité (l'environnement est disposé de telle sorte que les participants ne se voient pas par exemple), le participant pourra transmettre l'information par une autre modalité (vocale par exemple). L'utilisateur a ainsi le choix dans le cas (2) de fournir un backchannel par des gestes ou des vocalisations. Doter un agent conversationnel de la capacité à transmettre et interpréter des actions multimodales nécessite de fortes contraintes de synchronisation entre les actions entreprises sur les différentes modalités (Kopp *et al.*, 2006). Dans le cas (1), l'agent coordonne son discours avec une désignation gestuelle du lieu auquel il fait référence. S'il ne parvient pas à coordonner son geste et montre du doigt un lieu

différent du lieu auquel il fait référence dans son discours, il perdra le bénéfice de la production multimodale de son action et pire, il sèmera la confusion dans l'esprit de l'utilisateur. La temporalité des gestes co-verbaux est aussi d'une importance première : ces derniers doivent être synchronisés au bon moment dans la phrase. Cassel (2000) cite comme exemple la production d'un hochement de tête qui, selon l'endroit de l'énoncé où il est produit change sa signification.

2.4 Architectures informatiques d'agents conversationnels

2.4.1 Composantes d'une architecture d'agent

La capacité d'un agent à interpréter à la fois les signaux verbaux et non verbaux de l'utilisateur nécessite pour un agent de disposer de différentes composantes (Thórisson, 1999; Skantze et Hjalmarsson, 2010; Hartholt *et al.*, 2014). Le module de reconnaissance vocale est chargé de transcrire la parole de l'utilisateur en représentation textuelle. En parallèle différents capteurs de l'agent (suivi de gestes, oculomètre, par exemple) captent les signaux non-verbaux de l'utilisateur. La transcription textuelle de l'agent est interprétée par un module de reconnaissance de langue naturelle (*natural language understading* ou *NLU*) et les informations provenant des autres capteurs de l'agent par un module de reconnaissance des signaux non verbaux (*non-verbal behavior understanding* ou *NBU*) transformant les données verbales et non verbales en représentation sémantique exploitable par l'agent. L'agent peut alors traiter ces informations afin de raisonner sur l'action à réaliser selon l'information provenant de l'environnement et de son propre état cognitif ou émotionnel. Les intentions d'action de l'agent sont formulées en actes communicatifs transcrits ensuite en énoncés par un module de génération de langage naturel (*natural language generation* ou *NLG*) et actions non-verbales qui sont exécutées par un *behavior realizer*.

2.4.2 Premières architectures informatiques

La présence de ces composantes en elle-même ne suffit pas à implémenter un agent fonctionnel. Les problématiques de multimodalité, de gestion temps réel et de contraintes de synchronisation entre plusieurs processus de coordination parallèles énoncées section 2.3.2 nécessitent de définir la manière dont ces modules sont inter-connectés. Il existe un grand nombre d'architectures d'agents conversationnels créées dans les années 2000. Nous présentons ici trois exemples d'architectures informatiques, proposant trois philosophies différentes de conception.

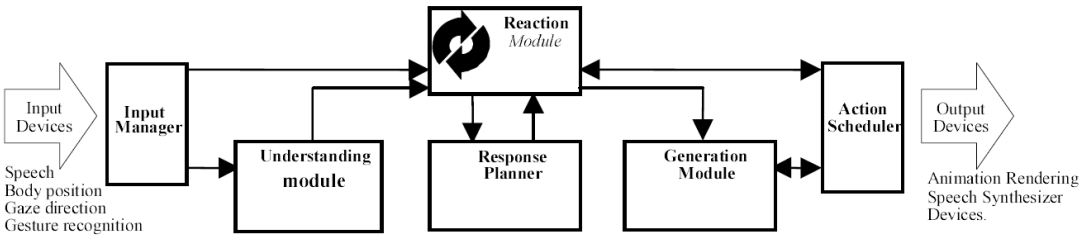


FIGURE 2.1 – Schéma représentant l'architecture REA. Extrait de Cassell *et al.* (1999).

Architecture STEVE de Rickel et Johnson (1997)

STEVE (Rickel et Johnson, 1997) s'inspire de l'architecture cognitive Soar (Laird *et al.*, 1987) pour implémenter un agent pédagogique utilisé pour l'apprentissage de tâches procédurales à des étudiants. Le comportement de STEVE est généré par l'exécution de plans d'actions hiérarchiques. Chaque étape du plan d'action comporte un lien causal définissant le but permettant la transition à l'étape suivante. Lors de la démonstration de tâches, STEVE garde en mémoire son intention à réaliser cette action et, si l'utilisateur le demande, l'agent est capable d'expliquer verbalement à l'utilisateur pourquoi il est en train de réaliser cette action. L'agent surveille aussi la progression de la tâche réalisée par les utilisateurs et met à jour son plan d'action en fonction des objectifs atteints ou non par les utilisateurs. L'architecture STEVE est divisée en deux modules, un module cognitif représentant l'intelligence de l'agent et un module sensori-moteur recevant les commandes d'action de l'agent et les exécutant.

Architecture REA de Cassell *et al.* (1999)

Cassell *et al.* (1999) ont conçu REA, une architecture capable, entre autres, de gérer l'échange de tours en situation dyadique d'interaction agent-utilisateur. L'architecture est capable de traiter des entrées provenant de plusieurs modalités et de générer à la fois des signaux verbaux et non-verbaux. Dans cette architecture, l'agent est capable :

- de fournir des signaux de *backchannels* lorsqu'il sent que l'utilisateur l'invite à en faire ;
- de gérer certains comportements de tour de parole ;
- de gérer les formules d'initiation (salutations par exemple) et de fin d'interaction.

L'organisation de l'architecture est illustrée sur la figure 2.1. Tel qu'illustrée sur cette figure REA comprend plusieurs modules.

1. Le gestionnaire d'entrée se charge de récupérer les entrées multimodales provenant des capteurs de l'agent. Plus particulièrement, l'architecture prend en compte la gestuelle, les événements audio et comporte une reconnaissance

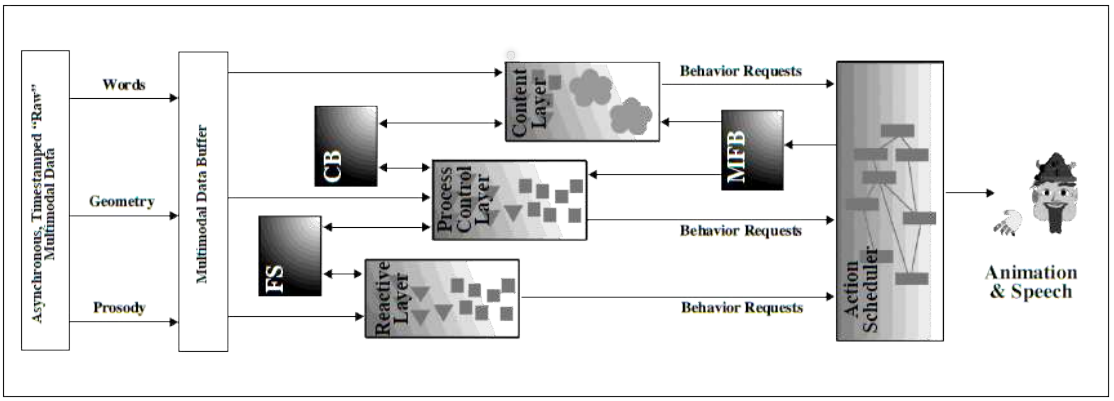


FIGURE 2.2 – Schéma de l'architecture Ymir. Les trois couches de traitement sont représentées ainsi que les *blackboards* permettant la communication entre modules. Un troisième *blackboard* le *motor feedback blackboard* contient des informations sur les actions en train d'être exécutées par l'agent, et sur la localisation des éléments de l'environnement. Extrait de Thórisson (1999).

vocale à base de grammaire ;

2. Le module de compréhension fusionne les modalités d'entrées (les signaux non-verbaux et ce que dit l'agent), met à jour l'état de la conversation et propose sur la base de cet état, les actions que l'agent doit réaliser ensuite ;
3. Le module de réaction est responsable de la sélection d'action de l'agent ;
4. Le module de planification de réponse formule des plans d'actions qui seront exécutés pendant les cycles ultérieurs de la simulation ;
5. Le module de génération d'action produit, à partir d'actions de haut niveau fournies par le module de réaction, des primitives d'action qui sont ensuite envoyées à l'ordonnanceur d'action ;
6. le module d'ordonnancement d'action est responsable du déclenchement des actions et assure la coordination entre les actions.

Dans cette architecture, le contrôle du comportement de l'agent est centralisé. La sélection de l'action de l'agent passe ainsi par le module de réaction qui est chargé de déterminer les actions à réaliser en fonction des données provenant de l'environnement.

Architecture Ymir de Thórisson (1999)

Ymir est une architecture d'agent conversationnel animé créée par Thórisson (1999), traitant de l'interprétation, de la génération d'actions multimodales et des contraintes temps-réel entre les modules. Contrairement à REA qui base le contrôle du comportement sur une approche centralisée, le postulat de base de Thórisson pour la construction de son architecture est que le comportement de l'agent émerge de l'interaction entre plusieurs processus cognitifs spécialisés. Le point clé de l'architecture est sa décomposition en sous-modules simples chacun spécialisés dans la

perception d'un type d'événement ou dans la réalisation d'une action. Ces modules s'exécutent indépendamment les uns des autres et la décision concernant l'action à exécuter résulte d'une compétition entre les modules. Deux types de modules s'exécutent dans l'architecture, les percepteurs se chargent de faire correspondre aux données provenant des capteurs de l'agent des données de perception plus abstraites et multimodales utilisées par les composants de prise de décision de l'agent. Les données produites par les percepteurs sont de nature booléennes. Ces percepteurs se divisent eux-mêmes en deux catégories, les percepteurs unimodaux s'occupent de la détection d'un signal multimodal provenant des capteurs de l'agent, tandis que les percepteurs multimodaux agrègent les données provenant d'autres percepteurs unimodaux et multimodaux. Les décideurs reçoivent un ensemble de données provenant des percepteurs, et se chargent d'exécuter une action selon les données reçues. Ces décideurs peuvent être « déclarés » (« overt »), déclenchant des requêtes d'action (« behavior requests ») envoyées ensuite au planificateur d'action (« action scheduler ») ou « non déclarés » (« covert »), provoquant un changement dans la perception de l'état du dialogue (par exemple les états « utilisateur engagé dans l'interaction » et « utilisateur auditeur ou locuteur »). Ce changement dans l'état du dialogue modifie en contre-partie la répartition des modules dans l'architecture. Certains modules spécialisés dans la perception des signaux du locuteur sont par exemple désactivés si l'agent passe à l'état auditeur et inversement. Lorsque le planificateur d'action reçoit les requêtes de comportement de la part des décideurs de l'architecture, il transforme ces requêtes en actions multimodales exécutées par l'agent, puis planifie le déclenchement de l'action selon les informations contenues dans la requête et la priorité de l'action.

Ces composantes sont réparties en plusieurs couches définissant plusieurs fréquences d'exécution des modules de l'architecture. L'utilisation de couches d'exécution permet de définir des temps de réaction et de mise en action précis pour l'agent. Trois couches sont implémentées dans l'architecture. La couche réactive, la plus prioritaire, concerne les comportements nécessitant un cycle de perception-action de moins d'une seconde. La couche de supervision du dialogue concerne les comportements ayant un cycle de perception-action de moins de deux secondes et représente ce qui est en rapport avec la gestion de la dynamique du dialogique (tour de parole, *grounding* par exemple). Les processus liés à la gestion du tour de parole ou au *grounding* sont implémentés dans cette couche d'exécution. La couche de traitement du contenu, la moins prioritaire, concerne les comportements ayant l'ordre de grandeur temporel le plus élevé et gère le traitement et la production d'informations relatives au contenu de la conversation. L'échange de données entre les modules se fait par l'intermédiaire de *blackboards*, partagés entre plusieurs couches d'exécution. Le *Functional Sketchboard* (« FS » sur la figure 2.2) est ainsi le *blackboard* permettant l'échange de données provenant de modules de perception de bas-niveau

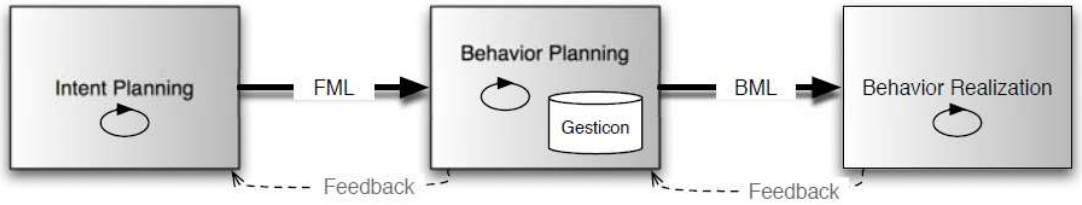


FIGURE 2.3 – Illustration de l'architecture SAIBA. Extrait de Kopp *et al.* (2006).

(mouvement, présence d'un signal vocal ou non, par exemple) entre les modules de la couche réactive et les modules de la couche de supervision du dialogue, tandis que le *Content Blackboard* (*CB* sur la figure) gère l'échange de données moins « critiques temporellement » tel qu'énoncé par Thórisson (1999). Un troisième *blackboard*, le *Motor Feedback Blackboard* (« MFC » sur la figure) stocke les informations liées à l'exécution des actions par l'agent. La figure 2.2, illustre cette répartition des modules en trois couches d'exécution, la communication entre les modules par l'intermédiaire de *blackboards* et la gestion du déclenchement des requêtes d'action par l'intermédiaire du gestionnaire d'action.

2.4.3 Standard SAIBA

Devant le constat que les premières architectures créées présentent des points communs dans leur implémentation, Kopp *et al.* (2006) proposent la définition d'un standard permettant de mutualiser les travaux réalisés par différents auteurs. En effet, la conception d'un agent conversationnel animé est une tâche complexe abordant différentes problématiques : acquisition des données provenant de l'utilisateur, mécanismes de prises de décision de l'agent, contrôle des actions en temps réel, rendu graphique de l'agent (Hartholt *et al.*, 2014). Ces problématiques sont vastes et difficilement abordables par une seule équipe de recherche (Hartholt *et al.*, 2014). De plus, la validation d'un module de comportement de l'agent par un laboratoire de recherche nécessite l'implémentation d'autres briques logicielles ne constituant pas le domaine de spécialité de ce laboratoire (Hartholt *et al.*, 2014). Le besoin de mutualisation et de réutilisation des travaux de différentes équipes de recherche s'est alors fait ressentir, résultant en un certain nombre de standards et architectures permettant la mutualisation des travaux.

SAIBA (Situation, Agent, Intention, Behavior, Animation) présentée par Kopp *et al.* (2006) constitue le premier effort en ce sens et aborde plus particulièrement la problématique de la génération de comportements multimodaux sans se soucier de celle de l'interprétation des données provenant de l'utilisateur. L'architecture proposée est décrite sur la figure 2.3. SAIBA divise la génération d'une action communicative en trois étapes assurées chacune par un module spécifique. L'*Intent Planner* reçoit en entrée les informations de perception provenant de l'utilisateur et génère les intentions communicatives verbales et non-verbales de l'agent (ce que va dire

l'agent ou prendre le tour par exemple). Le *Behavior Planner* reçoit alors les intentions communicatives de l'agent et sélectionne à partir de ces intentions les actions motrices à réaliser par l'agent. Le *Behavior Realizer* reçoit enfin les actions du *Behavior Planner* et se charge de lancer ces actions dans l'environnement 3D. Bien que le standard SAIBA renseigne le rôle de chaque module, il ne spécifie pas la manière dont les différents modules sont implémentés : cette tâche revient aux utilisateurs de l'architecture. Néanmoins, le standard propose deux langages pour la communication entre modules basés sur le langage XML (eXtensible Markup Language) : le FML (Function Markup Language) pour la communication entre l'*Intent Planner* et le *Behavior Planner* (Cafaro *et al.*, 2014) et le BML (Behavior Markup Language) pour la communication entre le *Behavior Planner* et le *Behavior Realizer* (Kopp *et al.*, 2006).

Le langage FML sert deux fonctions, la première est de renseigner l'état courant du dialogue par des informations concernant l'identité, la personnalité et le niveau de relation des personnes engagées dans les différentes interactions verbales, et des informations concernant les différentes interactions (appelé *floor*) dans lesquelles l'agent est engagé. Pour chaque *floor*, les rôles des participants (locuteur, interlocuteur, *side participant*) sont de plus renseignés. La seconde fonction est de spécifier les actes communicatifs de l'agent. Ces actes communicatifs sont spécifiés au niveau de l'interaction (gestion de l'engagement ou gestion du tour par exemple) ou au niveau des actes de langage (contenu de la phrase). Le FML spécifie aussi les états mentaux de l'agent pouvant influer sur la réalisation des comportements de l'agent.

Dans sa version courante, le BML spécifie, pour chaque action, la partie du corps réalisant l'action ainsi que le moment de démarrage et de fin de l'action. Ce moment de démarrage peut être renseigné par une valeur temporelle absolue ou relative : dans ce dernier cas, l'action démarre à un moment précis de la progression d'une autre action. Le niveau de détail concernant la description des modules et surtout la spécification du langage BML permet une réutilisabilité des composants de l'architecture. De même, les composants créés pour une architecture particulière fonctionnent pour toutes les autres architectures. Cette modularité de l'architecture encourage l'échange des différents composants entre utilisateurs de SAIBA.

SAIBA ne spécifie que la génération des actions communicatives multimodales des agents. Plus récemment Hartholt *et al.* (2014) ont proposé *Virtual Human Architecture*. *Virtual Human Architecture* propose une spécification des modules (appelés « capacités » dans l'architecture) nécessaires à la fois à la perception des comportements de l'utilisateur, mais aussi à la génération des comportements de l'agent. L'architecture est composée exactement de neuf capacités :

— la reconnaissance vocale, transformant la parole de l'utilisateur en représentation textuelle de ce que vient de dire l'utilisateur ;
— le NLU (*Natural Language Understanding*), transforme le texte fourni par

le module de reconnaissance vocale en information sémantique de plus haut niveau interprétée ensuite par l'agent ;
— la détection audio-visuelle, localisant et reconnaissant les expressions liées à la communication non-verbale (signaux visuels, prosodie,...) ;
— le module d'interprétation des comportements non-verbaux, combinant des informations de différentes modalités ;
— l'agent, comprenant un module de gestion du dialogue recevant la phrase prononcée par l'utilisateur et déterminant la réponse à fournir par l'agent et un module de génération d'intention communicative ;
— un module de génération de comportements non-verbaux recevant l'intention communicative de l'agent ;
— un module de génération de texte à partir de l'intention communicative reçue de l'agent ;
— un module de génération de parole recevant le texte à dire par l'agent, assurant la transcription en signal sonore et son exécution ;
— un module de réalisation de comportements, recevant les actions spécifiées par le module de génération de comportements non-verbaux et assurant la synchronisation de tous les comportements.

Conformément à SAIBA, *Virtual Human Architecture* (VHT) est modulaire et permet l'interchangeabilité de ses composants permettant l'échange de certains composants entre concepteurs d'agents conversationnels. L'architecture spécifie de même un système de communication entre les modules (VHMsg), et propose un ensemble de messages standards pour la communication basé en partie sur une spécification de FML et sur le langage BML.

Le standard SAIBA et l'architecture VHT souffrent de défauts pointés par d'autres auteurs de la littérature. Un des principaux manques identifiés est l'incapacité d'implémenter des comportements réactifs, incrémentaux ou continus. Dans SAIBA ou VHT, toute action réalisée par l'agent provient d'une intention formulée dans SAIBA par le langage FML. Bevacqua *et al.* (2009) montrent que certains comportements ne peuvent pas entièrement provenir d'une intention communicative fournie par le langage FML. Ainsi certains comportements comme le mimétisme dans une conversation sont parfois des comportements inconscients. Bevacqua *et al.* (2009) proposent ainsi que la modélisation de ces aspects de la communication ne devrait pas passer par l'*Intent Planner* mais devrait être gérée par un module de gestion du comportement réactif dans le *Behavior Planner*. Nooraei *et al.* (2014), mettent en avant la nécessité de concevoir des architectures suivant un modèle de "signalement continu", plus proches de ce que l'on peut observer dans les interactions humaines que les approches traditionnelles dont font partie les architectures SAIBA et VHT.

Dans ce modèle, l'agent et l'utilisateur interprètent et réagissent en continu aux actions mutuelles de chacun. L'agent n'attend pas la fin de l'action de l'utilisateur

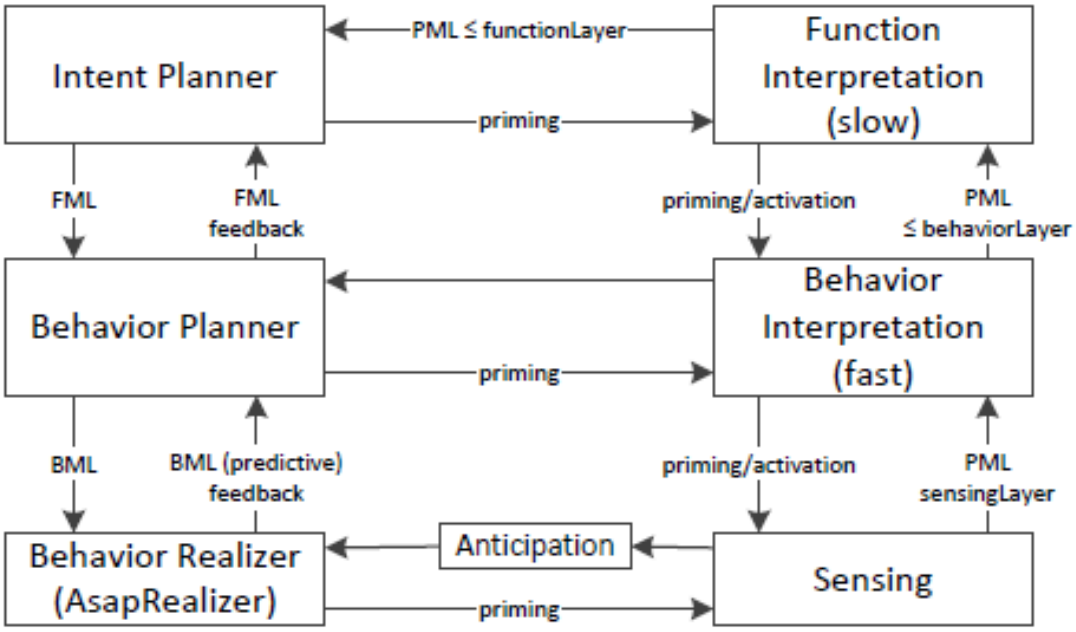


FIGURE 2.4 – Architecture ASAP. Extrait de Kopp *et al.* (2014).

pour commencer à l'interpréter, et potentiellement y réagir. Pendant l'exécution de son action, il surveille la réaction de l'utilisateur et module en continu ou interrompt l'action en cours selon l'action fournie par l'utilisateur. Ce modèle prend sa justification dans certains attributs des conversations humaines, notamment la capacité des auditeurs à produire des *feedbacks* coopératifs en simultané avec l'énoncé du locuteur courant, ce dernier surveillant ces *feedbacks* et pouvant modifier son énoncé s'il estime que son interlocuteur a mal compris ce qu'il disait (Clark, 1996).

Le modèle de signalement continu renvoie au fonctionnement des architectures incrémentales d'agent (Schlangen *et al.*, 2010; Kopp *et al.*, 2014; Skantze et Hjalmarsson, 2010). Nous présentons dans la section suivante ces architectures en présentant une des plus récentes, l'architecture ASAP (Kopp *et al.*, 2014).

2.4.4 Architecture pour la gestion incrémentale du dialogue : ASAP

L'architecture ASAP (*Artificial Social Agent Platform*) (Kopp *et al.*, 2014), schématisée sur la figure 2.4 permet l'implémentation de modules cognitifs incrémentaux. L'architecture est divisée en deux parties, une partie d'interprétation des signaux de l'utilisateur et une partie de génération de l'action. La partie de génération de l'action reprend les principes de l'architecture SAIBA (Kopp *et al.*, 2006).

ASAP étend le langage BML pour y ajouter des capacités incrémentales. Le langage BML étendu (ou *BMLa*) permet ainsi d'interrompre une action en cours et de la redémarrer plus tard ou de moduler les paramètres d'une action. Suivant l'architecture SAIBA, les modules du *Behavior Realizer* envoie des *feedbacks* sur le

statut de l'action en cours de réalisation. Les modules du *Behavior Planner* peuvent ainsi être informés de la progression des actions et réagir en cas d'interruption non voulue de l'action. Dans le cas d'un énoncé envoyé à la synthèse vocale, les modules de l'architecture peuvent ainsi suivre la progression de l'énoncé. Lorsque l'énoncé se termine ou est sur le point d'être terminé, le gestionnaire d'énoncés peut ainsi soit planifier la génération d'une autre énoncé, soit modifier l'intention communicative de sorte que l'agent laisse la parole à l'utilisateur.

La partie perception de l'action est elle-même composée de trois modules, le module de *Sensing* composé des différents capteurs de l'agent, le module de *Behavior Interpretation* interprétant à partir des données multimodales de l'agent, les différentes actions de l'utilisateur et le module de *Function Interpretation* interprétant la fonction communicative des actions de l'utilisateur. Les données sont échangées par le biais du langage PML (*Perception Markup Language*) (Zwiers *et al.*, 2011).

La perception et le contrôle de l'action, incrémentales dans l'architecture, reposent sur le modèle des unités incrémentales introduit par Schlangen et Skantze (2011). Dans ce modèle chaque action est divisée en une liste chainée d'unités incrémentales d'action. Les prédécesseurs de l'unité incrémentale représentent les unités d'action exécutées juste avant l'unité incrémentale actuelle et les successeurs de l'unité incrémentale représentent les unités d'actions exécutées juste après l'unité incrémentale. Une unité incrémentale peut posséder plusieurs successeurs. Ces successeurs représentent dans le cadre de la perception incrémentale de l'action plusieurs hypothèses alternatives sur le comportement en cours de l'agent. Les unités incrémentales peuvent avoir aussi une relation hiérarchique. Cette relation hiérarchique définit différents niveaux d'abstraction dans la relation entre les unités incrémentales. Par exemple, un énoncé prononcé par un utilisateur sera représenté au plus bas niveau d'abstraction par la liste des segments audio des phonèmes prononcés par les participants. À un niveau d'abstraction plus élevé la représentation textuelle des phonèmes sera spécifiée, puis les mots seront représentés, pour finir au plus haut niveau par la sémantique de l'énoncé. Lorsque l'agent interprète un énoncé prononcé par l'utilisateur il cherche dès la captation des premières unités incrémentales et sans attendre la récupération de tous les phonèmes à lier les segments audio avec les phonèmes prononcés puis avec les mots et enfin avec la représentation sémantique d'un ensemble d'unités incrémentales. À mesure que l'énoncé sera produit, il récupèrera de plus en plus d'unités incrémentales associées à la phrase, et sera capable de modifier les hypothèses sur ce que l'utilisateur essaie de dire à l'agent. La génération d'action incrémentale repose aussi sur ce principe. L'agent formulera au niveau d'abstraction le plus élevé la sémantique de l'acte de dialogue. Les différentes unités incrémentales seront indépendamment et immédiatement traduites en unités incrémentales à un niveau d'abstraction de moins en moins élevé jusqu'aux unités incrémentales qui seront exécutées par l'agent. La génération d'action incrémentale

permet la modification en temps réel d'une unité incrémentale qui n'a pas encore été exécutée par l'agent, même si les unités incrémentales précédant immédiatement cette dernière sont en train d'être exécutées. L'agent peut ainsi modifier son action à la volée.

La particularité de l'architecture ASAP est la possibilité de générer des actions sans passer par les modules traitant des actes communicatifs (*Intent Planner* et *Function Interpreter*). Les liens existants entre le module de *Behavior Interpretation* et le *Behavior Planner* représentent l'expression du contrôle de l'action de l'agent à un niveau non symbolique.

Chapitre 3

Modèles conceptuels pour le contrôle des échanges de parole utilisateurs-agents

3.1 Motivations

La problématique du tour de parole utilisateur-agent apparait dans le cadre d'échanges de parole non contraints par un dispositif physique, à l'opposé de dispositifs où l'utilisateur doit réaliser une action explicite comme appuyer sur un bouton pour signaler qu'il va parler (Balentine *et al.*, 1997), et par la nature de l'architecture proposée (voir la section 2.4.4). L'agent est ainsi libre de parler à n'importe quel moment dans la conversation. Pour éviter une interaction où l'agent interromprait systématiquement l'utilisateur, il est alors nécessaire de doter l'agent de capacités à savoir quand parler et quand laisser la parole. Les approches les plus simples traitant de cette problématique s'appuient sur un détecteur d'activité vocale pour détecter le début et la fin de parole de l'utilisateur. L'agent peut alors raisonner sur des règles simples telles que « ne prendre la parole que si l'utilisateur ne parle pas » et « s'interrompre lorsque l'utilisateur se met à parler » pour gérer les échanges de paroles. Néanmoins, et au vu de la manière dont la gestion des tours de parole s'effectue dans les interactions humaines, cela génère deux problèmes principaux : d'une part l'agent est incapable de distinguer les pauses des fins de tour et d'autre part de distinguer entre des *backchannels* ou des paroles de l'utilisateur non destinées au système et des tentatives d'interruption. Cette problématique a donné lieu au développement de modèles de gestion du tour de parole plus complexes.

3.2 Modèles existants

3.2.1 Distinction entre pauses et fins de tour dans des interactions dyadiques

Les approches les plus simples résolvant la problématique de la distinction entre pause et fin de tour sont des approches par temporisations. L'idée de ces approches est d'attendre un certain seuil temporel avant de considérer que l'utilisateur a fini son tour (Ward *et al.*, 2005). Néanmoins le seuil idéal de temporisation utilisé pour éviter à l'agent de prendre le tour lors d'une pause de l'utilisateur conduit à des interactions hachées, peu fluides et peu naturelles (Ward *et al.*, 2005). Certains auteurs se sont penchés sur la détection de signaux indiquant à l'agent que l'utilisateur est en train d'effectuer une pause plutôt qu'une fin de tour. Cassell *et al.* (1999) proposent ainsi un algorithme simple pour détecter une pause en interprétant la gestuelle de l'utilisateur. Lorsque l'utilisateur s'arrête de parler, le système détecte une pause si l'utilisateur continue à produire des gestes pendant le moment de silence et une fin de tour autrement. Skantze *et al.* (2014) présentent, eux, un système contrôlant la direction du regard d'un robot permettant de mieux signaler à l'utilisateur des pauses du robot par rapport à des fins de tour.

Néanmoins, la majorité des approches s'intéressent, elles, à l'interprétation de signaux de fin de tour pour permettre l'optimisation de la transition de tour en évitant toute interruption accidentelle de l'agent liée à la mauvaise interprétation d'une pause comme une fin de tour. Plusieurs auteurs ont ainsi choisi d'exploiter un certain nombre de signaux produits par l'utilisateur et corrélés aux fins de tour pour détecter de manière fiable la fin de tour de l'utilisateur et réduire les temps de transition. Ces approches peuvent être divisées en plusieurs catégories. Les approches utilisant des règles expertes et manuellement codées, les approches probabilistes et les approches utilisant de l'apprentissage automatique pour s'adapter au locuteur en face de lui.

En ce qui concerne les approches sur la base de règles expertes, Thórisson (2002) propose une architecture informatique et un modèle conceptuel (*Ymir Turn-Taking Model* ou *YTTM*) pour les échanges de parole basé sur l'architecture *Ymir* (voir page 40). L'auteur propose dans cette approche un certain nombre d'hypothèses inspirées de la littérature pour l'implémentation de la gestion du tour de parole. L'ensemble de ces hypothèses constituent ce que Thórisson (2002) appelle un modèle génératif d'échange de parole : il décrit les processus cognitifs en jeu dans la gestion du tour de parole plus que les simples observations effectuées dans des interactions humaines. Le modèle a été implémenté dans l'architecture informatique Ymir (Thórisson, 1999). L'implémentation effective des hypothèses de YTTM donne lieu à un système de règles tel qu'illustré sur la figure 3.1.

Dans ce modèle, la détection de la fin de tour est produite par un système de

Other-is-giving-turn **11**
ACTIVE-DURING-STATE: *Other-Has-Turn*
CONDITIONS:
(AND
(Other-is-speaking = F)
(OR
(AND
(Other-is-looking-at-me = T)
(Other-is-facing-me = T))
(AND
(Other-is-looking-at-me = T)
(Other-is-gesturing = F))
(AND
(Other-is-gesturing = F)
(Other-is-facing-me = T))))

Other-accepts-turn **16**
ACTIVE-DURING-STATE: *I-Give-Turn*
CONDITIONS:
(AND
(Other-is-looking-at-me = F)[c]
(Other-is-presenting = T))

Other-is-addressing-me **17**
CONDITIONS:
(AND
(Other-is-turned-to-me = T)
(Other-is-facing-me = T)
(Other-is-looking-at-me = T))

FIGURE 3.1 – Exemple de règles implémentées dans l'architecture Gandalf de Thorisson. Extrait de Thórisson (2002).

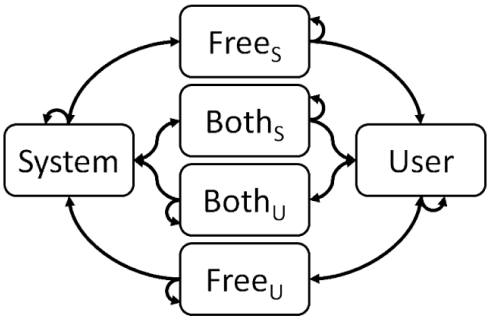


FIGURE 3.2 – Illustration de la machine à états utilisée par Raux et Eskenazi (2012). Extrait de Raux et Eskenazi (2012).

règles prenant en compte la direction de la tête et du regard, la gestuelle pour déterminer si un utilisateur laisse le tour.

Huang *et al.* (2011) appliquent un mécanisme d'apprentissage automatique pour entrainer un agent à détecter des fins de tour dans des enregistrements d'interactions humaines à partir de la direction du regard, de la prosodie, des hochements de tête du locuteur courant et des indices syntaxiques de l'énoncé de ce dernier. Ils proposent comme approche de validation du modèle la comparaison entre les performances du modèle dans la précision (capacité à détecter la fin de tour au moment où elle a lieu) et l'exactitude (discriminer des fins de tour de pause) de la détection des fins de tour provenant des enregistrements d'interactions humaines et les performances de participants humains pour la même tâche. Ils montrent que le modèle est capable d'atteindre les performances humaines en termes de détection pour environ 50 % des fins de tour.

L'approche prise par Raux et Eskenazi (2012) est un modèle probabiliste pour la gestion du tour de parole dans le cadre d'une interaction dyadique avec l'utilisateur. L'agent surveille au cours du temps l'état du dialogue et assigne à chacun de ces états une probabilité de se trouver dans cet état à un instant donné. La machine

à états utilisée par Raux et Eskenazi (2012) est illustrée sur la figure 3.2. L'état *System* correspond aux situations où le système a le tour, l'état *User* correspond à la situation où l'utilisateur a le tour. La transition ne se fait pas directement de l'utilisateur à agent mais passe par des états intermédiaires $Free_s$, et $Free_u$ pour indiquer la fin de tour respectivement de l'utilisateur et de l'agent. Les états $Both_s$ et $Both_u$ sont des états correspondant aux interruptions du locuteur courant par, respectivement, le système et l'utilisateur. Les auteurs entrainent le modèle à trouver les valeurs des signaux de l'utilisateur permettant de discriminer une fin de tour d'une pause provenant d'un corpus d'interaction humain-agent. Les signaux utilisés ici comprennent la structure du discours de l'utilisateur (question fermée, question ouverte, ...), la sémantique de l'énoncé de l'utilisateur, le volume sonore des dernières syllabes, la hauteur de voix finale et la durée de la dernière syllabe. Les auteurs montrent la capacité de leur modèle entrainé à diminuer la valeur moyenne de durées de transition à 500 ms et le taux moyen de recouvrements à 3 %.

Jonsdottir et Thórisson (2013) proposent une architecture où la gestion du tour de parole avec l'utilisateur est apprise au cours de l'interaction avec ce dernier. Pour cela les auteurs s'inspirent de l'architecture YTTM pour implémenter un module de gestion du tour de parole sur la base unique des variations de prosodie des participants. Un mécanisme d'apprentissage par renforcement est utilisé pour sélectionner la meilleure action à réaliser en fonction des attributs prosodiques de l'agent. Les auteurs testent l'efficacité de la prise de parole de l'agent dans le cadre d'échanges verbaux avec dix utilisateurs. L'agent est capable, pour chaque interaction avec l'utilisateur, de coordonner la prise de parole avec l'utilisateur afin de reproduire les durées de transition observées dans des interactions humaines.

3.2.2 Distinction de paroles coopératives et compétitives

Peu d'approches ont abordé la problématique de la distinction entre les interruptions de l'utilisateur et divers énoncés coopératifs (sans prise de tour) ou paroles de l'utilisateur non destinées au système (des paroles destinées à d'autres personnes ou des pensées à voix haute de l'utilisateur). La capacité de l'utilisateur à interrompre le système est intéressante car elle mène potentiellement à des dialogues plus efficaces entre utilisateurs et agents. L'utilisateur n'est ainsi plus obligé d'attendre que l'agent finisse de parler pour pouvoir prendre la parole mais peut couper l'agent dès qu'il a compris ce que ce dernier voulait dire. Néanmoins, autoriser des paroles simultanées avec le système génèrerait des faux positifs dans la reconnaissance de la parole de l'utilisateur. L'agent pourrait en effet s'arrêter de parler alors que l'utilisateur ne lui parlait pas ou mal reconnaitre la parole de l'utilisateur du fait du bruit généré par la phrase prononcée par l'agent (Selfridge *et al.*, 2013; Witt, 2014). En vue de discriminer des énoncés coopératifs sans prise de tour d'une tentative d'interruption, les résultats de Reidsma *et al.* (2011) ont montré la difficulté à distinguer en temps

réel un tel énoncé d'une interruption (voir section 4.3). Pour pallier cette difficulté, plusieurs auteurs ont proposé d'autres approches pour diminuer le nombre d'interruptions inappropriées du système. Selfridge *et al.* (2013) proposent un algorithme destiné aux systèmes incrémentaux permettant d'évaluer la probabilité que la phrase de l'utilisateur soit comprise par l'agent. Si la phrase est comprise par le système, ce dernier se met en pause, au contraire si la phrase n'est pas comprise par le système celui-ci continue de parler. Witt (2014) propose un modèle estimant la probabilité de prise de parole de l'utilisateur à un instant donné. Le modèle est bâti à partir de données statistiques provenant d'interactions utilisateur-système. Cette probabilité pondère la détection de parole de l'utilisateur. Une détection de parole de la part du détecteur d'activité vocale (chargé de déterminer si l'utilisateur est en train de parler ou non) peut ainsi être considérée comme un faux positif par évaluation de cette probabilité. À l'inverse, si la probabilité que l'utilisateur parle à un instant t est forte et si le volume sonore capté est proche mais inférieure au seuil de détection d'activité vocale, le système pourra considérer que l'utilisateur parle à cet instant donné.

3.2.3 Tour de parole multipartite

Les modèles ci-dessus suivent le principe de minimisation des durées de transition introduit par Sacks *et al.* (1974) : les participants à une conversation s'échangent la parole de sorte de minimiser les durées de transitions tout en évitant la détection de fin de tour. Ce principe de minimisation des transitions de tour entre utilisateur et agent est une problématique générale des architectures de gestion des tours de parole. Néanmoins, lorsque l'on s'intéresse à des conversations multipartites les différents rôles que prennent les auditeurs (Clark, 1996) dans une conversation posent le problème de la désignation de l'interlocuteur et du prochain locuteur. Cette désignation est réalisée en grande partie par les regards des participants. Les études de Ishii *et al.* (2006) et de Johansson *et al.* (2014) montrent ainsi la pertinence d'implémenter la gestion du regard pour réguler la conversation dans un cadre multi-partie. Al Moubayed et Lehman (2015) évaluent l'impact du mouvement des yeux et de la tête séparément dans la régulation du tour de parole entre l'agent et plusieurs utilisateurs enfants. Ils observent que le mouvement des yeux seuls n'est pas suffisant pour désigner le prochain locuteur, mais le mouvement de la tête en complément des yeux est un signal beaucoup plus perçu par les utilisateurs.

Bohus et Horvitz (2011) proposent une approche probabiliste optimisant un seuil de temporisation avant de prendre la parole. L'objectif de ce travail est de permettre à un agent de minimiser son temps de prise de tour tout en évitant des conflits de parole liés à l'auto-sélection des participants.

3.2.4 Prise en compte des intentions de l'agent

Ravenet *et al.* (2015) proposent un modèle où l'attitude de l'agent influence différentes dimensions du comportement d'un agent, notamment ses stratégies de prises de parole. Dans leur modèle, le statut et l'affiliation de l'agent jouent un rôle modulateur dans la variation du temps de prise de parole, à partir d'une valeur seuil de dominance ou de convivialité, l'agent décide de prendre la parole en interrompant le locuteur précédent. Lessmann *et al.* (2004) et Thórisson *et al.* (2010) suivent un principe similaire, ils intègrent dans leur architecture d'agent une « intention de parler » (Lessmann *et al.*, 2004) et une « urgence à parler » (Thórisson *et al.*, 2010). Lessmann *et al.* (2004) modélisent « l'intention de parler », donnant la capacité à l'agent de prendre la parole de manière pro-active en produisant des signaux de prise de tour et décider de continuer à parler ou non si l'utilisateur l'interrompt. L'« urgence à parler » (Thórisson *et al.*, 2010) est, elle, une variable continue ayant plusieurs dimensions : la probabilité qu'un autre participant veuille parler, la vitesse à laquelle l'urgence à parler monte (une forme d'impatience à parler), et la tolérance à laisser le tour à quelqu'un d'autre si ce dernier veut parler aussi. Thórisson *et al.* (2010) proposent une implémentation d'un modèle de tour de parole étendant YTTM (Thórisson, 2002) pour la modélisation d'échanges multipartites, incluant cette variable d'urgence à parler. Une simulation dans le cadre d'une conversation entre douze agents montre la capacité de son modèle à faire varier le nombre de recouvrements et la durée totale des silences dans la conversation en modulant cette variable.

Au contraire, l'urgence à parler, à la fois pour l'utilisateur et pour l'agent, peut être nulle à un instant donné générant un moment de silence gênant. La présence de silences trop longs sont problématiques dans le cadre d'interactions utilisateur-agent, et peuvent rapidement provoquer un désengagement de l'utilisateur (Ohshima *et al.*, 2015). Ohshima *et al.* (2015) proposent ainsi un agent capable de générer un *filler* (voir glossaire) lors de moments de silence gênant pour montrer à l'utilisateur son intention de continuer et pousser l'utilisateur à recommencer à parler.

3.3 Positionnement

Une grande majorité des modèles de gestion du tour de parole sont créées dans le but d'assurer le flux continu de la conversation en évitant des erreurs de détection d'une fin de tour de la part de l'agent. Ces architectures se basent en grande majorité sur le principe de minimisation des silences : les participants s'échangent la parole afin de minimiser les silences en évitant les recouvrements de parole au cours d'une conversation (Jonsdottir *et al.*, 2008; Bohus et Horvitz, 2011; Raux et Eskenazi, 2012). Ils adhèrent ainsi à l'hypothèse énoncée par Cutler et Pearson (1986) qu'une interruption prématurée de la conversation peut être systématiquement remontée

à une erreur dans les échanges de parole. Une grande majorité des évaluations de modèle ont ainsi été effectuées en mesurant à quel point l'agent était capable de minimiser ses temps de prises de parole et de diminuer le taux de recouvrement avec la fin du tour de l'utilisateur (Huang *et al.*, 2011; Raux et Eskenazi, 2012; Jonsdottir et Thórisson, 2013).

Les approches mentionnées ci-dessus suivent le paradigme d'un agent poli, au service de l'utilisateur. Au contraire, l'engagement de l'utilisateur pourrait être amélioré s'il percevait que l'agent avait ses propres buts indépendants de ceux de l'utilisateur. Nous rejoignons ainsi les positionnements de Lessmann *et al.* (2004), Thórisson *et al.* (2010) et Ravenet *et al.* (2015) modulant les stratégies de résolution de conflits et de prise de parole selon une « intention de parler » (Lessmann *et al.*, 2004) de l'agent. La prise en compte de ces intentions est peu intéressante pour des interactions d'initiative non-mixte (interactions de type question-réponse) entre l'utilisateur et l'agent, l'interaction étant déjà structurée afin d'imposer un ordre préétabli dans l'échange de tour. Cependant, pour aller vers des conversations utilisateur-agent, au sens d'interactions informelles et d'initiatives mixtes entre les participants, la modulation des stratégies de prise de parole et de résolution de conflits est essentielle pour prendre en compte tous les scénarios liés aux échanges de parole (O'Connell *et al.*, 1990). Nous sommes ainsi particulièrement intéressés par la gestion de situations de conflit ou de silences gênants dans le dialogue en plus d'assurer un échange fluide de la parole de l'agent. Jonsdottir et Thórisson (2013) ne nient pas l'intérêt de ce genre de modèle mais le considèrent comme un mécanisme à part entière, distinct de la gestion du tour de parole. Par un seul modèle de coordination des échanges de parole nous souhaitons au contraire rendre compte de toutes les situations liées à la gestion des tours de parole humains. Dans ce modèle, les situations de conflits, de silences gênants ou de transitions courtes sont partiellement pilotées par les intentions de parler des agents. Une intention forte de parler des deux participants résulterait en un conflit, tandis qu'une intention de parler faible des participants résulterait en un silence gênant, potentiellement géré par la production de *filler* (Ohshima *et al.*, 2015).

Chapitre 4

Tour de parole dans les conversations humaines

4.1 Définition d'un tour

Dans la majorité des conversations, on observe que les échanges verbaux entre les participants sont structurés séquentiellement. Les prises de paroles des participants sont organisées de sorte qu'un participant parle à la fois (Sacks *et al.*, 1974). L'intervalle temporel où un participant parle est considéré alors comme un tour de parole selon Bernstein (1962). Néanmoins, l'échange de tours n'est pas parfaitement séquentiel et une conversation est parsemée de silences entre les tours, de pauses dans le discours du locuteur et de moments de recouvrements entre les participants (Heldner et Edlund, 2010). Ces recouvrements de parole sont en partie dus à la présence de *backchannels* parlés produits par l'auditeur pendant que le locuteur courant parle. Ces moments de parole courts n'apportent aucun nouvel élément à la conversation mais servent à un interlocuteur à signaler son statut d'auditeur attentif dans la conversation. Ces occurrences ne semblent pas pour autant problématiques et constitueraient des termes ou vocalisations utilisées par un ou plusieurs auditeurs pour donner « licence » selon les termes de Clancy et McCarthy (2015) à un locuteur courant à continuer la production de son énoncé. Les *backchannels* sont ainsi rarement considérés comme des tours de parole (Goodwin, 1981). De même l'occurrence de silences et de paroles simultanées est une problématique en soi lorsque l'on veut caractériser les échanges de parole (Goodwin, 1981). Considère-t-on un moment de la conversation où deux personnes parlent comme deux tours de parole simultanés ou comme un seul ? Attribue-t-on les silences entre les tours comme appartenant au locuteur courant ou au locuteur suivant ? Est-ce que deux paroles successives d'un même locuteur séparées par du silence doivent être considérées comme deux tours différents ou comme un seul tour (Goodwin, 1981) ? Une des raisons à cette difficulté provient du fait qu'on ne peut définir un tour qu'en regard du comportement de chaque participant. Ainsi, Yngve (1970) considère que personne n'a autorité

pour définir quel participant a le tour à un instant donné, et que l'idée d'un tour elle-même est intrinsèquement liée aux croyances que chaque participant a sur la personne qui a le tour à un instant donné plus que sur la production de parole des participants. Un tour de parole n'existerait alors à un moment de la conversation que parce que les participants ont choisit conjointement de désigner à ce moment de la conversation un seul locuteur possible (Sacks *et al.*, 1974). Cela complexifie la définition d'un tour, les segments de silence et de parole ne suffisant pas à délimiter un tour de parole. Définir et étudier le tour de parole revient donc à découvrir comment les participants décident de l'attribution d'un tour (Sacks *et al.*, 1974; Duncan, 1972).

4.2 Modèles généraux d'étude du tour de parole

Deux grandes approches expliquent et définissent le processus de gestion du tour de parole dans une conversation : l'approche par réaction aux signaux et l'approche par projection.

4.2.1 Approche par réaction

La première, l'approche par réaction aux signaux (Duncan, 1972), définit la gestion des tours de parole comme un mécanisme servant à réguler le rythme de la conversation en empêchant « les déviations des comportements appropriés ». Le locuteur est défini comme le participant qui revendique le tour de parole tandis que l'auditeur est un participant qui ne revendique pas le tour de parole. Aussi, lorsqu'un auditeur se met à revendiquer le tour sans accord du locuteur, on observe un état de tours simultanés. Ce sont ces états de tours simultanés qui constituent les déviations des comportements appropriés, et, lorsque ceux-ci apparaissent, cela constitue une rupture du mécanisme de gestion du tour de parole. Les participants résolvent les états de tours simultanés en suivant un ensemble de règles considérées comme indépendantes du mécanisme de coordination de la parole. Pour éviter ces états de tours simultanés, un ensemble de signaux de prise de tour (Duncan et Niederehe, 1974) et de signaux de fin de tour (Duncan, 1972) existent basés sur une combinaison variable d'indices comportementaux verbaux et non verbaux utilisés par les participants pour se coordonner. Les signaux sont ici de nature événementielle, un participant produit un signal qui est interprété dans sa totalité (baisse ou augmentation de hauteur de voix à partir de 300 ms avant le moment de silence, la terminaison ou l'initiation d'un geste par exemple) par ses interlocuteurs qui lui répondent par l'initiation d'une fin de tour ou par une prise de tour. Néanmoins, la production de ces signaux n'impose pas un changement de tour, lorsque l'auditeur produit des signaux de prise de tour, le locuteur peut choisir d'abandonner ou non le tour, de même, l'auditeur peut choisir de prendre le tour ou non à la suite des

signaux de fin de tour du locuteur courant. En ce sens, les changements de tour sont négociés par les participants (Duncan, 1972).

4.2.2 Approche par projection

L'approche par projection (Sacks *et al.*, 1974) propose une alternative à l'explication des échanges de tour par emploi de signaux de fin de tour et de début de tour. Sacks *et al.* (1974) observent que la majorité des transitions de tour se font sans moments de silence : les participants prennent le tour à la suite du locuteur courant sans avoir attendu la terminaison de la phrase. Pour expliquer cette observation, les auteurs postulent que les tours d'un locuteur sont composés d'unités de construction de tour (*turn constructional unit* ou *TCU*), des phrases entières, expressions ou mots uniques permettant aux auditeurs de prédire le contenu verbal produit après cette unité et identifier les points de terminaison sémantique des participants. Ces points de terminaison constituent des « moments pertinents de transition » (*transition relevant place* ou *TRP*) représentant des opportunités pour les auditeurs de prendre le tour. Cette capacité à anticiper les *TRP* permettrait aux participants de préparer leur énoncé et de prévoir précisément le moment où ils commenceront à parler afin de minimiser la transition de parole. Aux *TRP*, les transitions de tour se font selon des règles précises.

1. (a) Si le locuteur courant sélectionne explicitement quelqu'un, celui-ci est obligé de prendre le prochain tour, les autres ne peuvent prendre le tour.

   (b) Si aucun futur locuteur n'est sélectionné, l'auto-sélection peut mais pas obligatoirement être mise en place, le premier qui prend la parole acquiert le droit de prendre le tour et le transfert se passe à ce moment.

   (c) Si le tour du locuteur courant est construit de manière à ne pas sélectionner le prochain locuteur alors le locuteur courant peut mais n'est pas obligé de continuer sauf si un autre auditeur s'auto-sélectionne.

2. Si le locuteur courant continue de parler à la suite de ce *TRP*, l'ensemble des règles ci-dessus est appliqué de nouveau au *TRP* suivant.

Pour chaque participant l'ensemble des règles s'exécute ici de manière séquentielle, la règle 1(a) est d'abord appliquée puis la règle 1(b) et la règle 1(c). L'application de la règle 1(b) laisse la possibilité à plusieurs participants de s'auto-sélectionner pour le tour suivant, provoquant un conflit de parole résolu en ne laissant qu'un participant prendre définitivement la parole. La pression des autres participants pousse donc un participant souhaitant s'auto-sélectionner à le faire le plus rapidement possible pour s'assurer d'être le prochain possesseur du tour, expliquant selon Sacks *et al.* (1974) les transitions de parole courtes observées au cours de la conversation. Ces règles amènent Sacks *et al.* (1974) à définir trois caractéristiques fondamentales du tour de parole.

1. Le mécanisme est laissé au contrôle des participants. La gestion s'effectue soit par le locuteur courant sélectionnant le locuteur suivant, soit par les locuteurs suivant s'auto-sélectionnant et les autres participants prenant ou laissant le tour.

2. Cette administration se fait par l'interaction, la contribution de chaque participant à la détermination de l'ordre des tours dépend des contributions des autres participants. Le locuteur emploie au cours de son tour des *TCU* en vue de faciliter la projection de son tour par les auditeurs. Cependant, un *TRP* ne conditionne pas obligatoirement la prise de tour d'un potentiel locuteur suivant, et peut résulter en un locuteur courant continuant sont tour. C'est la prise de parole d'un participant, ce dernier ayant projetté l'opportunité de prendre le tour par l'interprétation des *TCU*, qui conditionne l'arrêt d'un tour. De même la prise de tour d'un participant est acceptée par le silence des autres participants.

3. La gestion est locale, le système n'alloue qu'un tour à la fois, et ne prédétermine pas à l'avance l'organisation ni la taille des tours. Chaque transition de tour ne sélectionne que le locuteur du tour suivant immédiatement la transition. De même la durée du tour dépend de l'application des règles 1(a) et 1(b) de sélection d'un participant lors de l'occurrence d'un *TRP* et ne peut donc être prédite par le locuteur au début de sa prise de parole.

Les auteurs traitent, dans leur approche, les recouvrements de parole liés à une lutte pour la prise de tour comme des événements apparaissant aux *TRP* et liées à un conflit entre deux participants s'étant auto-sélectionnés ou, dans le cas d'une tentative de prise de tour en dehors d'un *TRP* comme une « violation du système de gestion du tour de parole ». Dans les deux cas, ces recouvrements sont résolus par un mécanisme autre que le mécanisme de gestion du tour de parole (Schegloff, 2000).

4.2.3 Validité des approches

Les deux approches présentées ci-dessus ont majoritairement été utilisées pour étudier les tours de parole dans les conversations. Il existe néanmoins un débat quant à la validité de ces approches. Les partisans de l'approche par réaction aux signaux ont souligné le manque de preuves étayant l'affirmation de Sacks *et al.* (1974) concernant l'optimalité des transitions de tour (Heldner et Edlund, 2010). En réponse à ces critiques, plusieurs auteurs ont mesuré les durées de transition. Schegloff (2000) observe par exemple que « la durée normale de transition de tour » est de l'ordre de la durée d'une syllabe c'est-à-dire entre 150 et 200 ms. Certaines occurrences de transition de tour étant même inférieures à 100 ms (Thórisson, 2002). Les mêmes mesures pour la langue française montrent une durée médiane de transition de tour

de 451 ms (Campione et Véronis, 2002) soit l'équivalent de deux à trois syllabes prononcées par un participant. Ces durées moyennes de silence sont inférieures au temps d'interprétation de l'énoncé du locuteur courant puis de planification nécessaire pour commencer à produire une phrase, la durée de planification se situant autour de 1500 ms pour des phrases simples (Torreira *et al.*, 2015). Cela nécessiterait que le participant soit capable de projeter en avance la fin du tour. Cela semble plaider en faveur de l'approche par projection. Or, cela est vrai si l'on considère que la planification de la phrase à prononcer ne se fait que lorsque l'auditeur a identifié une fin de tour probable. Au contraire, des études récentes comme celle de Bögels *et al.* (2015) montrent que la planification de la phrase se ferait bien avant de repérer la fin du tour. L'auditeur, ayant déjà l'énoncé qu'il va prononcer en tête, n'aurait pas besoin d'autant de temps pour se mettre en action, laissant la possibilité d'une réaction aux signaux de fin de tour. Heldner et Edlund (2010) argumentent ainsi en considérant que le temps de réaction minimum d'un participant pour détecter la fin de tour du locuteur courant et prendre le tour est de l'ordre de 200 ms. Pour évaluer si les transitions sont projetées ou non, Heldner et Edlund (2010) mesurent les durées de transition provenant d'interactions entre participants anglophones. Cela constitue, pour eux, le critère permettant de trancher en faveur d'une projection ou d'une réaction, les durées de silences inférieures à 200 ms ne pouvant être réalisé selon eux que par une projection et les durées de silences supérieures à 200 ms étant réalisées par une réaction aux signaux de fins de tour et de prise de tour. Ainsi les recouvrements représentent dans leur étude 40 % des transitions et 40 % des transitions sont des moments de silence de plus de 200 ms. Ils concluent en affirmant que les participants utilisent à la fois la projection et la réaction aux signaux pour prendre le tour.

L'expérimentation de Grosjean et Hirt (1996) remet en cause la capacité d'un participant, dans certaines situations, à anticiper la fin de tour du locuteur courant. Dans le protocole expérimental qu'ils proposent, les participants écoutent trois types de phrases en anglais et en français. La première, la plus simple, se termine par un verbe et un groupe nominal (par exemple « Earlier my sister took a dip »), les deux autres types de phrases commençant de la même manière mais se finissant par des propositions optionnelles. Une des phrases ajoute trois mots après le nom (« Earlier my sister took a dip in the pool ») et l'autre ajoute six mots (« Earlier my sister took a dip in the pool at the club »). Chaque phrase est segmentée en huit morceaux de sorte que chaque morceau de phrase contient un segment de plus que le morceau de phrase précédent (d'abord « Ear » puis « Earlier » jusqu'à « Earlier my sister took a dip »). Les participants sont divisés en quatre groupes, chaque groupe écoutant deux morceaux différents sélectionnés aléatoirement. À chaque morceau entendu les participants renseignent dans un questionnaire le type de phrase qu'ils pensent entendre (« Earlier my sister took a dip », « Earlier my sister took a dip in the pool »,

« Earlier my sister took a dip in the pool at the club »). Les résultats montrent une difficulté des participants à distinguer le type de phrase qu'ils écoutent sauf au dernier mot terminant la partie commune des trois phrases. Lorsque les participants écoutent le groupe nominal terminant la partie commune des trois phrases (« a dip »), ils sont capables de distinguer si la phrase finit après ce mot ou continue, mais lorsque la phrase continue, ils sont incapables de prédire le nombre de mots restants. Cette expérimentation montre que lorsque des énoncés comportent des propositions optionnelles, il est impossible pour un participant de prédire quand la phrase va se terminer. Il ne peut alors se fier qu'à la hauteur de voix finale selon Grosjean et Hirt (1996) pour prédire la fin de la phrase.

Pour savoir si le processus en jeu dans l'estimation de la fin de tour est bien un processus de prédiction, Magyari et de Ruiter (2012) proposent une étude expérimentale où les participants écoutent un ensemble de fins de tour provenant d'interactions humaines en néerlandais, puis sont chargés de deviner le nombre de mots terminant la phrase. Les auteurs observent un lien entre une mauvaise perception de la fin de tour et une incapacité à estimer le nombre de mots restants. Lorsque les participants détectent la fin de tour après son occurrence réelle ils anticipent aussi un nombre de mots supérieur au nombre de mots prévu. À l'inverse, lorsque les participants estiment la fin de tour trop tôt par rapport à l'occurrence réelle de la fin de tour, ils sous-estiment le nombre de mots restants dans le tour du locuteur. Les auteurs observent néanmoins une difficulté des participants à anticiper le nombre de mots avant la fin du tour, avec une prédiction correcte du nombre de mots en fin de tour inférieure à 50 % des réponses des participants.

Riest *et al.* (2015) cherchent à évaluer si les participants estiment la fin de tour par prédiction ou par réaction. Les participants sont soumis à trois conditions expérimentales où ils sont chargés de réagir le plus rapidement possible à la fin de tour en écoutant des enregistrements en néerlandais de fins de tour provenant d'interactions humaines. Dans l'une des conditions, ils connaissent à l'avance ce que va dire le locuteur courant, dans une autre condition, ils ne connaissent pas l'énoncé du locuteur, dans la dernière condition, l'enregistrement est altéré, rendant toute information linguistique inaudible par les participant, la seule manière de savoir si le locuteur courant a fini le tour est de réagir à des signaux prosodiques potentiels de fin de tour (ils ne connaissent pas non plus à l'avance le contenu de l'énoncé). Les auteurs n'observent pas de différences significatives dans le temps de réaction des participants selon que ces derniers connaissent la phrase à l'avance ou non. Ils observent une distinction significative entre ces deux conditions et la condition où l'information linguistique est enlevée de l'enregistrement. L'observation d'un temps de réaction similaire pour les deux premières conditions, amène les auteurs à conclure que les participants dans la seconde condition, possèdent les mêmes informations que les participants ayant une connaissance préalable de l'énoncé dans la première condi-

tion. Ces informations seraient alors acquises par projection de la fin de tour du locuteur courant.

En comparaison, plusieurs études expérimentales montrent des corrélations entre certains signaux de fin de tour potentiels et les fins de tour effectives des participants. Ces signaux sont de nature para-verbale (Gravano et Hirschberg, 2011), verbales (Gravano et Hirschberg, 2011), ou non-vocales (Novick *et al.*, 1996; Holler et Kendrick, 2015). Bien que des corrélations existent, cela n'implique pas que ces variations systématiques soient exploitées par les participants pour détecter la fin de tour du locuteur. Cela ne résout donc pas la problématique de l'utilisation de ces signaux. Pour résoudre cette question, Niebuhr *et al.* (2013) présentent une expérimentation de perception utilisateur suivant le même principe que le protocole de Grosjean et Hirt (1996). Les participants écoutent un ensemble de phrases en allemand, possédant un début similaire mais avec une partie optionnelle ou non selon la condition. Pour chaque ensemble de phrases, les participants sont confrontés à des énoncés pour lesquels le profil de volume sonore et de hauteur de voix du dernier phonème de la partie commune sont modifiés entre les conditions. Les auteurs observent que selon ces variations de signaux prosodiques, les participants sont capables d'anticiper plus ou moins exactement si la phrase se termine juste après le phonème écouté ou si celle-ci continue, montrant l'impact de variations du volume sonore et de la hauteur de voix sur la perception de la fin de tour.

4.2.4 Critique des modèles de Duncan (1972) et Sacks *et al.* (1974)

Il est difficile de trancher en faveur de l'une ou l'autre des approches. Aucun protocole expérimental n'a permis de favoriser l'approche par projection par rapport à l'approche par réaction. De plus ces approches introduisent chacune des problématiques non résolues actuellement. Ces problématiques poussent certains auteurs à critiquer ces deux approches traditionnelles.

O'Connell *et al.* (1990) mettent en avant le caractère contextuel des échanges de parole. Ils rejettent l'idée que le seul critère de succès d'une conversation soit un échange de tour où un participant parle à la fois. Lors de discussions animées, les participants parlent simultanément, se coupent la parole à chaque tour, et pourtant cela ne constitue pas en soi un échec de la conversation puisque les attentes des participants par rapport à la conversation sont comblées (O'Connell *et al.*, 1990). Aussi, d'autres facteurs contextuels doivent être pris en compte pour comprendre la manière dont les participants s'échangent la parole dans une conversation.

Clark (1996) critique l'existence de règles que les participants suivent explicitement pour s'échanger la parole. Selon eux, les paroles alternées sont une conséquence du processus de *grounding*. Les participants ont d'une part la nécessité de s'assurer que la phrase soit prononcée dans les conditions optimisant sa compréhension et

d'autre part fournissent une organisation permettant la facilitation des échanges de signaux de *grounding*. Les contributions verbales sont ainsi divisées en une phase de présentation et une phase d'acceptation. Lors de la phase de présentation le locuteur produit un acte communicatif. Une fois la phase de présentation effectuée, le locuteur requiert des indices positifs de la part de l'interlocuteur concernant la compréhension de l'acte communicatif du locuteur. Celui-ci le fait en signalant sa compréhension ou non de l'énoncé du locuteur dans la phase d'acceptation. C'est aussi, selon Clark (1996), cette alternance des phases de présentation et d'acceptation qui explique la variabilité des transitions de tour que l'on peut observer pendant la conversation. L'interlocuteur peut ainsi choisir d'interrompre le locuteur courant et prendre la parole plus tôt s'il pense avoir déjà compris ce que voulait dire le locuteur courant. Au contraire, une incertitude le conduira à prendre la parole avec un moment de silence plus long. Clark (1996) met en avant l'existence de terminaisons collaboratives, des types particuliers de contributions où l'interlocuteur offre une terminaison possible à un énoncé non terminé. Cela s'observe particulièrement lorsque le locuteur courant ne trouve pas les mots pour finir la phase de présentation, dans ce cas, l'interlocuteur propose à sa place une fin d'énoncé. Ces terminaisons collaboratives sont particulièrement étudiées par Clancy et McCarthy (2015) qui avancent que ce type de terminaison montre que les énoncés sont co-construits au fil de l'interaction par tous les participants, locuteur et interlocuteur, rendant difficilement prédictible la fin de tour du locuteur courant. De même, les *backchannels* conditionnent la progression du tour du locuteur (Clancy et McCarthy, 2015).

Plusieurs études mettent en avant le caractère contextuel des échanges de parole. Kilpatrick (1986) observe que 95 % des tours commencent ou finissent par des paroles simultanées dans le cas d'échanges entre des participants portoricains. De même Berry (1994) montre des différences dans la manière dont les recouvrements compétitifs sont résolus entre une conversation entre quatre participants américains et une conversation entre quatre participants espagnols. Les recouvrements durent ainsi deux fois plus longtemps dans les conversations entre les participants espagnols par rapport aux conversations entre les participants américains.

Ter Maat *et al.* (2010) montrent qu'une différence de perception des attitudes de l'interlocuteur peut être induite chez les participants en variant les moments de prise de parole de parole d'un agent. Ils proposent une étude expérimentale où un participant interagit avec un agent conversationnel. Ce dernier varie ses stratégies de prise de parole entre les conditions de sorte que, pour une condition, il attendra un moment de silence avant de prendre la parole alors que, pour une autre condition, il cherchera à prendre le tour juste avant la fin de tour de l'auditeur. Ils observent que les participants jugent l'agent plus poli mais peu assuré lorsqu'il laisse un moment de silence entre les tours, et plus agressif et plus assuré lorsqu'il prend la parole juste avant la fin de tour de l'utilisateur. Ces résultats, obtenus dans le cadre d'interactions

entre utilisateur et agent, laissent penser qu'une variation des stratégies de prise de parole aurait le même effet dans les interactions humaines,

En complément de ces critiques, des modèles alternatifs d'échanges de tours ont été formulés. Ces modèles s'intéressent aux processus cognitifs sous-jacents dans la gestion des échanges de parole. Wilson et Wilson (2005) se basent sur un constat réalisé lors de l'observation d'interactions dialogiques humaines : les moments de silence dans les transitions de tour ne sont pas répartis uniformément entre les participants, mais semblent être multiples d'une unité de temps (Wilson et Zimmerman, 1986; Bailly et Gouvernayre, 2012). De même, plusieurs processus d'alignement, notamment dans les cycles de respiration des participants à l'approche de la fin de tour ont été observés McFarland (2001). Sur la base de ces constats, les auteurs émettent l'hypothèse que les transitions de tour résulteraient de l'alignement d'oscillateurs cérébraux endogènes entre les participants. Les oscillateurs impliqués dans les échanges de parole seraient perçus par la détection du cycle de prononciation des syllabes. Lorsque le locuteur parle, il ne peut initier une nouvelle syllabe que lorsque le potentiel de l'oscillateur est à son maximum. Au contraire, au milieu de la prononciation d'une syllabe, le potentiel est au plus bas : le locuteur ne peut initier de nouvelles syllabes. L'auditeur détecte la fréquence d'oscillation du locuteur par le biais de ce cycle de prononciation, et synchronise son oscillateur en opposition de phase avec l'oscillateur du locuteur. De la même manière que le locuteur, l'auditeur ne pourra prendre le tour que si son oscillateur est à son potentiel maximum, c'est-à-dire, à des moments précis par rapport à la fin du tour du locuteur.

Ikegami et Iizuka (2007) montrent pour un agent artificiel comment un « style » de changement de tour peut émerger du couplage sensorimoteur continu d'une dyade, sans utilisation de règles explicites de changement de tour, ni de connaissance à priori de « signaux » de changement de tour entre les participants. Ils proposent un modèle d'interaction entre deux agents se déplaçant dans un espace en deux dimensions. Pour cette interaction, les deux agents sont engagés dans une tâche de poursuite où, alternativement, un agent est meneur, et l'autre agent suit la trajectoire de l'agent meneur. De temps en temps, les rôles entre les participants changent, l'un passe de meneur à suiveur et inversement. L'objectif des auteurs dans cette tâche est d'obtenir les transitions de rôle les plus courtes possibles. Pour atteindre cet objectif, les auteurs font évoluer les réseaux de neurones suivant un algorithme génétique. La sélection est réalisée en évaluant la capacité des agents à prédire correctement au cours de la simulation les mouvements de l'autre et la capacité de l'agent, lorsqu'il est suiveur de suivre l'autre agent en minimisant la distance et l'angle entre lui et son partenaire. Au fil des générations d'agents obtenus, l'agent devient de plus en plus autonome vis à vis de son partenaire et ses trajectoires deviennent de moins en moins systématique. Les auteurs observent alors des agents capables de changer de rôle à des agents provenant de différentes générations, en variant leurs trajectoires lors des

changements de rôles selon le type d'agent. Ces différentes trajectoires n'étant pas pré-codées dans l'agent sont émergentes de l'interaction entre l'agent et l'utilisateur.

4.3 Variations de signaux observées dans les conversations

Au-delà des différentes approches prises pour modéliser les processus cognitifs et règles liés aux échanges de paroles, des corrélations entre différents signaux et différentes situations communicatives ont été observées.

4.3.1 Fins de tour

La question des ressources utilisées par le locuteur courant pour signaler sa volonté de laisser le tour aux autres participants est une question majeure de la littérature sur le tour de parole. Sur la base de l'observation de conversations face-à-face, Duncan (1972) identifie six signaux utilisés par les participants pour transmettre le tour : une variation de hauteur de voix positive ou négative à la fin d'un phonème, une voix « trainante » sur la dernière syllabe, la terminaison de n'importe quel geste de la main, des expressions stéréotypées, une baisse de la valeur de hauteur de voix et du volume sonore lorsque le locuteur prononce une expression stéréotypée et la terminaison d'une clause grammaticale. Les six signaux trouvés par Duncan (1972) ne sont pas systématiquement corrélés aux fins de tour. Ainsi le locuteur a le choix, selon l'auteur, des signaux qu'il souhaite produire pour finir son tour. Néanmoins, Duncan (1972) montre que les signaux s'additionnent : plus le nombre de signaux produits par les participants en fin de tour est grand, plus il y a de chances d'observer une transition de tour entre les participants.

Plusieurs auteurs ont, à l'instar de Sacks *et al.* (1974), identifié différents signaux corrélés à la fin de tour des participants. Les indicateurs de fins de tour les plus étudiés et les plus débattus sont les signaux para-verbaux. Gravano et Hirschberg (2011) explorent ainsi l'emploi de plusieurs signaux pour signaler la fin de tour dans le cadre d'échanges où les participants ne se voient pas. Confirmant les résultats de Duncan (1972), ils identifient des variations de signaux para-verbaux systématiquement produits avant les fins de tour des participants. On en compte quatre : des variations de hauteur de voix négatives ou positives sur le dernier phonème, une baisse plus générale de cette même hauteur de voix pendant la dernière seconde de la fin de tour, une augmentation de caractéristiques liées à la qualité de la voix (rapport signal sur bruit, *jitter* et *shimmer*) dans la dernière seconde, une baisse d'intensité et une diminution de la vitesse de parole (calculée par le nombre de phonèmes par seconde) sur le dernier mot. Cependant, bien que des corrélations soient trouvées, cela n'implique pas que ces signaux soient réellement exploités par

les participants avant une fin de tour. Pour adresser la question de l'exploitation ou non de ces signaux, Hjalmarsson (2011) propose une expérimentation de perception dans lequel des participants, ayant comme langue maternelle le Suédois, écoutent une conversation en suédois entre deux participants humains. De temps en temps l'extrait est mis en pause et l'utilisateur est chargé de déterminer si le locuteur courant continue de parler ou si l'auditeur prend le tour après la pause. Hjalmarsson (2011) montre ainsi la contribution d'une variation de hauteur de voix négative dans la perception de la fin de tour.

Plusieurs auteurs débattent, eux, de l'influence des terminaisons para-verbales en comparaison de celle des terminaisons lexico-syntaxiques pour la prédiction présumée de la fin de tour. De Ruiter *et al.* (2006) considèrent que la hauteur de voix pourrait être inutile pour la perception de la fin de tour, mais servirait plutôt de mécanisme de secours lorsque la projection de la fin de tour n'est pas possible. De Ruiter *et al.* (2006) proposent une expérimentation où les participants sont confrontés à des enregistrements audio de fins de tour provenant de conversations humaines. Il est demandé aux participants d'appuyer le plus rapidement possible lorsqu'ils jugent que le locuteur a fini son tour. Les auteurs observent une capacité des participants à estimer de manière exacte la fin de tour des participants 186 ms avant la fin réelle du segment de parole. Ces résultats reflètent, pour les auteurs, la capacité des participants à projeter la fin de tour. Afin de déterminer quels types d'indices sont utilisés pour projeter la fin de tour, l'expérimentation est divisée en cinq conditions, une condition *NATURAL* où l'utilisateur écoute l'enregistrement original de la conversation, une condition *NO-PITCH* où les variations de hauteur de voix sont enlevées de l'enregistrement, une condition *NO-WORDS* où toute information lexicale est enlevée de l'enregistrement, une condition *NO-PITCH-NO-WORDS* où ne sont gardées que les informations relatives à la variation de volume et une condition *NOISE* où aucune information autre que la présence ou non de parole ne peut être distinguée. Les résultats ne montrent pas de différence dans l'estimation de la fin de tour entre les conditions *NATURAL* et *NO-PITCH* mais montrent une dégradation de l'estimation lorsque l'information lexicale est enlevée avec une estimation de la fin de tour en moyenne 500 ms avant la fin de tour pour la condition *NO-WORDS* et de 700 ms pour la condition *NO-PITCH-NO-WORDS* (sur des enregistrements en moyenne de 3 s). Les durées moyennes d'anticipation pour ces deux dernières conditions sont néanmoins significativement plus proches de la condition naturelle que de la condition *NOISE*. De Ruiter *et al.* (2006) concluent que la nature des informations utilisées par les participants pour anticiper la fin de tour sont de nature syntaxique et sémantique, la hauteur de voix n'étant pas utile pour la projection, sauf pour discriminer des pauses de fins de tour. Au contraire, selon Bögels et Torreira (2015), la hauteur de voix joue un rôle critique dans la perception de la fin de tour. Dans une tâche où les participants sont chargés de répondre à un ensemble de questions,

ils montrent qu'aucun participant ne répond aux points de terminaison syntaxiques en absence d'un profil de hauteur de voix marquant, en complément, la fin du tour. La hauteur de voix servirait ainsi à « discriminer entre des points de terminaison de tour potentiels de points de terminaison réels de tour »(Lammertink *et al.*, 2015). Néanmoins, indicateurs verbaux et de hauteur de voix semblent complémentaires dans leur utilisation. Lammertink *et al.* (2015) montrent ainsi que les participants n'anticipent pas la fin de tour du locuteur courant lorsque celle-ci est syntaxiquement incomplète, mais réagissent aux variations de hauteur de voix et de volume sonore. Hauteur de voix et syntaxe semblent même interchangeables. Casillas *et al.* (2015) et Keitel et Daum (2015) montrent en ce sens que des nourrissons d'un an parviennent à anticiper les fins de tour, ces derniers ne pouvant le faire que par l'interprétation des signaux prosodiques. En comparaison avec ces résultats, dans l'étude de Keitel et Daum (2015), le retrait des informations liées à la hauteur de voix ne dégrade pas la capacité de participants adultes à percevoir les fins de tour semblant montrer que l'information de hauteur de voix semble superflue pour un adulte.

Concernant les informations syntaxiques liées aux fins de tour autres que les terminaisons syntaxiques, des expressions stéréotypées comme la présence de *tag questions* (questions ajoutées en fin de tour, équivalents pour la langue anglaise de l'expression « n'est-ce-pas ? »), sont observées dans les fins de tour des participants (Sacks *et al.*, 1974).

En ce qui concerne les informations provenant de signaux visuels, l'indicateur le plus étudié semble être le regard des participants. Kendon (1967) observe que les locuteurs ne regardent fixement leur interlocuteur qu'à la fin de leur tour. Novick *et al.* (1996) confirment ces observations en montrant des corrélations entre les regards échangés et les échanges de tour. L'étude de ces derniers montre que les variations de regards entre les participants ne semblent pas le fait d'un seul participant : on observe une coordination entre locuteur et auditeur. Deux modes de coordinations sont observées lors des fins de tour :

— *mutual-break* : le locuteur finissant son tour regarde vers l'auditeur courant, s'ensuit une période de regard mutuel avant que l'auditeur courant brise le regard mutuel et commence à parler, représentant 42 % des échanges de tour observés dans les conversation,
— *mutual-hold* : le type de coordination est quasiment identique au *mutual-break*, excepté que l'auditeur courant commence à parler en maintenant le regard mutuel, représentant 29 % des échanges de tour.

Stivers *et al.* (2009) mettent en évidence l'influence du regard du locuteur courant sur le temps de réponse du locuteur suivant, plus court lorsque le locuteur courant regarde le locuteur suivant. Le regard a aussi un rôle clé dans le cadre d'interactions multi-parties où il constitue un moyen de désigner un prochain locuteur sans le

nommer (Holler et Kendrick, 2015).

4.3.2 Initiation d'un tour

Dans l'approche de Sacks *et al.* (1974), les seuls indices utilisés pour la gestion des tours sont les *TCU* fournis par le locuteur permettant aux auditeurs d'anticiper un *TRP*. Dans cette vision, l'auditeur interprète passivement ces signaux sans avoir la possibilité d'indiquer aux autres participants sa volonté de prendre le tour. De plus, la nature du débat entre l'approche par projection et l'approche par réaction a amené la majorité des auteurs à se concentrer sur d'éventuels indicateurs de fin de tour et sur la capacité des participants à les interpréter tels quels. Quelques études se sont néanmoins intéressées à la manière dont un auditeur pouvait signaler son intention de parler. Indiquer sa volonté de prendre le tour ou non est particulièrement nécessaire dans le cas de *feedbacks* produits en parole simultanée avec le locuteur. En effet, un *feedback* produit par un participant peut être produit de manière compétitive (résultant en une tentative d'interruption) ou coopérative (conservant la répartition des rôles actuels) (Reidsma *et al.*, 2011). Pour Duncan et Niederehe (1974), la capacité d'un locuteur à distinguer entre une prise de tour d'un participant et un *feedback* parlé coopératif (un *feedback* parlé coopératif n'implique pas de volonté de prendre le tour) proviendrait de sa capacité à exploiter des signaux fournis par l'auditeur avant sa prise de parole. Ils distinguent quatre signaux principaux de prise de tour : l'initiation d'une gesticulation, l'emploi d'inspirations bruyantes, le détournement de la direction du regard et une intensité sonore plus forte. Reidsma *et al.* (2011) explorent la capacité d'un algorithme statistique à apprendre à distinguer un *feedback* produit de manière coopérative d'une prise de tour. Il observent que les caractéristiques les plus utilisées par l'algorithme pour distinguer ces deux situations après la durée des recouvrements sont acoustiques : la hauteur de voix, le volume sonore, le flux spectral et la qualité de la voix. Néanmoins ces résultats montrent que la latence moyenne de discrimination entre ces situations varie entre 300 ms et 1.1 s avec un taux d'erreur de 33 %. Pour aider à discriminer ces deux types de contribution de l'auditeur plusieurs auteurs proposent d'analyser le contexte dans lequel est réalisé le *feedback*. Gravano et Hirschberg (2011) mettent en évidence des signaux d'invitations aux *backchannels* : des variations d'indicateurs non-verbaux et des indicateurs verbaux du locuteur courant précédant un *backchannel* par l'auditeur. Ils classent comme signaux d'invitation au *backchannel* : une hauteur de voix finale montante et supérieure à la hauteur de voix moyenne du participant, une valeur d'intensité supérieure à la moyenne du participant, un ratio signal sur bruit supérieur et une structure grammaticale particulière (un énoncé se finissant par « déterminant-nom », « adjectif-nom » ou « nom-nom »). Ces signaux d'invitation aux *backchannels* seraient intentionnellement produits par le locuteur et constitueraient une raison pour laquelle ce dernier est capable de discriminer entre

une prise de tour et un *backchannel*.

Une étude similaire a été réalisée par de Kok (2013) pour des conversations en hollandais. Dans cette étude, de Kok (2013) met en évidence une corrélation entre une baisse de volume sonore, une baisse ou une augmentation de hauteur de voix et le taux de regards dirigés vers l'auditeur et la fréquence des *feedbacks* de l'auditeur. De plus, lorsque le locuteur emploie ces trois signaux, la probabilité d'avoir un *feedback* de l'auditeur est plus grande.

Un certain nombre de signaux permettent au locuteur suivant de signaler sa prise de tour au locuteur courant tout en lui permettant de retarder la production de l'énoncé. Cela sert par exemple lorsque l'énoncé du participant est difficile à formuler et qu'il a besoin de temps supplémentaire pour le planifier (Torreira *et al.*, 2015; Ohshima *et al.*, 2015). Certains de ces signaux sont des expressions verbales stéréotypées (de type « euh » en français) (Fox Tree, 2000), des *fillers* (Ohshima *et al.*, 2015) ou des inspirations bruyantes (Torreira *et al.*, 2015).

4.3.3 Garder le tour et résolution des conflits

Une autre catégorie de signaux est utilisée par le locuteur courant pour signifier sa volonté de garder le tour. Ces signaux sont employés lorsque le locuteur courant effectue une pause dans la progression de son tour afin d'empêcher l'utilisateur de prendre le tour pendant le moment de la pause. Duncan (1972) met en évidence les gesticulations, terme générique désignant toute forme de mouvement de la main et des bras sauf des gestes d'auto-contact, comme signal discriminant une pause, d'une fin de tour. Il définit le terme « gesticulation » comme toute forme de production gestuelle autre que des gestes d'adaptation (se gratter l'arrière du crane ou jouer avec un stylo pour signaler son ennui par exemple). Duncan (1972) observe que lorsque le locuteur effectue une pause dans son discours pour réfléchir à ce qu'il va dire, il aura tendance à continuer à produire des gesticulations, ce qui n'est pas le cas pour une fin de tour. Skantze *et al.* (2014) montrent qu'un détournement de regard peut servir à empêcher l'interlocuteur de prendre le tour. Les *filler* constituent de même des signaux utilisés par les participants pour signaler leur pause (Ohshima *et al.*, 2015; Clark et Fox Tree, 2002).

Une dernière catégorie de signaux est utilisée par les participants comme ressources pour résoudre les conflits de parole. Schegloff (2000) montre que lors de l'occurrence de recouvrements simultanés entre les participants, le volume sonore augmente, la hauteur de voix augmente et la vitesse de parole est modifiée (ralentie ou accélérée). Kurtić *et al.* (2013) observent que lorsque le locuteur courant est en conflit avec l'auditeur cherchant à parler, il a une tendance à répéter la dernière syllabe précédant le conflit de parole. Ce processus est appelé le recyclage. Le recyclage constitue l'indicateur discriminant le plus un conflit d'un recouvrement non compétitif avec une augmentation de la hauteur de voix et du volume sonore selon

Kurtić *et al.* (2013).

4.4 Positionnement

Les études sur le tour de parole présentées ci-dessus montrent indéniablement l'existence d'une coordination fine des échanges de paroles entre les participants. S'il n'existait pas de tels processus, on n'observerait pas de transitions d'une durée inférieure au temps de réaction nécessaire pour détecter une fin de tour et prendre la parole (Heldner et Edlund, 2010; Stivers *et al.*, 2009). Les participants attendraient simplement le silence après la fin de tour pour prendre le tour ou prendraient la parole sans tenir compte de l'autre. Pour expliquer ces coordinations, deux modèles principaux des processus liés à la gestion des tour de paroles des participants ont été élaborés et utilisés pour analyser les prises de parole. L'approche par réaction (Duncan, 1972; Duncan et Niederehe, 1974) postule que les échanges de parole sont coordonnés par le biais de signaux explicitement fournis par le locuteur courant et suivant. L'approche par projection (Sacks *et al.*, 1974) postule que les auditeurs sont capables de connaitre en avance le moment où le tour va se finir sur la base d'indices lexicaux et para-verbaux. Ils sont ainsi capables d'optimiser leurs échanges de paroles de sorte de minimiser les temps de transition. Ces deux approches postulent de plus que la gestion du tour de parole est un « mécanisme »fondé sur des règles explicites suivies par les participants pour éviter une « cassure » dans le cours normal de la conversation (Sacks *et al.*, 1974). Néanmoins aucune étude expérimentale n'a montré que l'un des deux processus constituait la manière réelle dont les participants s'échangeaient la parole. Bien que certaines études montrent notamment des corrélations entre des variations de signaux non-verbaux et les transitions de tour entre participants (Gravano et Hirschberg, 2011; Novick *et al.*, 1996; Bavelas *et al.*, 1995), elles ne montrent pas que celles-ci sont réellement exploitées par les participants pour coordonner leurs fins de tour. Au contraire, l'étude de Niebuhr *et al.* (2013) montre que le volume sonore et la hauteur de voix sont des variations de signaux activement exploitées par les participants pour distinguer une fin de phrase d'une fin de tour. L'interprétation de ces signaux est considérée par d'autres auteurs comme anecdotique et inutile au vu de la capacité des participants à détecter les fins de tour de manière précise uniquement sur la base du contenu verbal de l'énoncé (De Ruiter *et al.*, 2006). Les signaux para-verbaux sont vus au mieux comme des signaux de secours au cas où les participants sont incapables de projeter la fin de tour sur la base du contenu. En effet, au regard des durées de transition, inférieures au temps normalement nécessaire pour réagir au silence et prendre la parole, la réaction à des occurrences de signaux semble peu plausible. Pour rendre compte de la variabilité des transitions, (Heldner et Edlund, 2010) émettent l'hypothèse que certaines fins de tour sont anticipées et d'autres sont identifiées par l'interprétation de signaux

de fins de tour. Néanmoins, la projection elle-même n'est qu'une hypothèse qui n'a jamais pu être vérifiée expérimentalement. Au contraire, les faibles performances dans la prédiction du nombre de mots avant la fin de tour dans l'expérimentation de Magyari et de Ruiter (2012) et la problématique des tours constitués de clauses optionnelles (Grosjean et Hirt, 1996) montrent des failles dans l'explication de la détection de la fin de tour par anticipation du contenu sémantique. Il semblerait que les participants n'aient pas besoin de ce type de prédiction pour identifier la fin de tour du participant.

Contrairement à ce que les auteurs affirment, l'expérimentation de Riest *et al.* (2015) ne montre pas nécessairement que la capacité à prédire aussi bien la fin de tour peu importe la connaissance préalable ou non de l'énoncé implique que le processus en jeu est une prédiction du contenu lexical avant la fin de tour. Si les participants n'exploitent pas cette information pour coordonner leurs fins de tour, connaitre ou non le contenu sémantique ne changerait rien non plus au résultat. De plus ces mesures de la capacité à prédire la fin de tour du locuteur ont été faites sur des énoncés pré-enregistrés. Le caractère dynamique, co-construit des énoncés (Clancy et McCarthy, 2015) dans le cadre de conversations réelles rendrait plus difficile la prédiction de la fin de tour. L'existence de « règles » universelles à tout types de conversations, et l'idée que les participants ont pour objectif explicite de coordonner en permanence leurs prises de parole de sorte d'éviter les silences et les recouvrements de parole parait aussi contredire certaines observations réalisées sur les variabilités des situations liées aux échanges de parole (O'Connell *et al.*, 1990; Clark, 1996; Berry, 1994). Si ni la projection, ni l'interprétation de signaux événementiels sont utilisées dans le cadre de la gestion du tour de parole, quel processus est utilisé par les participants pour coordonner leurs échanges de parole ?

Le modèle génératif de Wilson et Wilson (2005) propose une approche novatrice pour la gestion des tours de parole. Cette approche se base sur la détection du cycle de prononciation des syllabes des participants en percevant les accents toniques dans l'énoncé prononcé par le locuteur. Néanmoins, bien que ces accents soient bien détectables pour certaines langues, ceux-ci sont moins évidents à percevoir pour d'autres langues (Cummins, 2012). De plus la notion de syllabe étant un concept du langage écrit, la correspondance entre syllabes et phonèmes dans la langue orale n'est pas toujours évidente à réaliser (Cummins, 2012). Une approche entièrement dynamique et dénuée de règles comme l'approche d'Ikegami et Iizuka (2007) est très intéressante à explorer, elle simplifie grandement l'explication de la manière dont une forme de prise de parole peut émerger sans règles préétablies, sans mécanisme explicite de coordination entre les participants. Cette approche est éloignée néanmoins des problématiques liées au tour de parole, elle propose une approche générique où les participants changent de rôle entre meneur et suiveur dans une tâche sans lien avec le tour de parole et dans le cadre d'une simulation entre

deux agents. Au vu du manque de modèle satisfaisant pour la gestion du tour de parole, nous proposons dans le chapitre suivant d'explorer plus les modèles traitant de manière plus générale des interactions humaines.

Chapitre 5

Processus cognitifs pour la coordination de la parole dans les interactions humaines

Nous souhaitons proposer un modèle pour la coordination du tour de parole entre un agent et un utilisateur qui permette, non seulement l'échange effective de tours entre les participants, mais soit motivé par des modèles de psychologie cognitive. La conversation est une tâche collaborative (Sacks *et al.*, 1974; Clark, 1996) où les participants coopèrent pour satisfaire un objectif commun. Nous explorons ainsi, en premier lieu, différents modèles de la manière dont les participants gèrent ces activités collaboratives, puis nous nous focalisons plus particulièrement sur la manière dont les participants coordonnent leurs actions dans le cadre de ces activités. Les modèles cognitifs de la collaboration reposent sur des modèles généraux de psychologie cognitive expliquant les processus de perception, de décision et d'action au sein d'un collectif d'humains collaborant pour accomplir une tâche commune. Nous explorons ensuite les théories liées à la collaboration et la coopération entre humains. Nous justifions enfin sur la base de cet état de l'art l'utilisation d'un paradigme peu utilisé dans le cadre de la gestion du tour de parole et des agents conversationnels animés, celui de la dynamique comportementale théorisé par Warren (2006).

5.1 Tâches collaboratives, coopération et coordination

5.1.1 Collaboration

Nous nous focalisons ici sur une définition de la collaboration, utilisée principalement dans la conception de systèmes informatiques favorisant le travail collaboratif entre utilisateurs (Ellis *et al.*, 1991).

La résolution de certaines tâches ne peut être réalisée par un individu seul.

Dans ce cadre, la formation de collectifs offre de nouvelles possibilités d'action et permet l'accomplissement des objectifs individuels de chaque participant (Marsh *et al.*, 2006). Par exemple, lorsqu'un individu doit saisir un objet, il peut, selon ses capacités motrices, prendre cet objet à une main ou à deux mains (Frank *et al.*, 2009). Lorsque les deux mains de l'individu ne suffisent plus à prendre l'objet, celui-ci sollicite un autre participant pour accomplir sa tâche (Fowler *et al.*, 2008). Apparait alors une activité collaborative. La collaboration est définie comme une activité où des individus travaillent ensemble sur un problème commun. Pour Ellis *et al.* (1991), la collaboration suit la règle des 3C, divisant celle-ci entre trois processus fondamentaux.

1. La communication : impliquant l'échange de messages (la conversation par exemple mais pas seulement) entre les participants pour mettre à jour leur connaissance mutuelle sur l'état de l'environnement et la progression de la tâche à accomplir.
2. La coopération : le fait d'opérer avec les autres dans un environnement partagé.
3. La coordination : l'ensemble des actions et processus permettant aux participants de faciliter leur communication et leur coopération, permettant particulièrement de traiter les conflits dans la réalisation de leur tâche (Pimentel *et al.*, 2004). Les systèmes d'alternance de tour, non seulement dans le cadre de la conversation mais aussi pour alterner l'accès à une ressource exclusive (Sacks *et al.*, 1974), sont des exemples de processus de coordination.

Lors d'une interaction, les participants doivent donc s'échanger de l'information (communiquer), agir ensemble (coopération) et synchroniser leurs actions (coordination). La figure 5.1 montre que communication, coopération et coordination sont interconnectées (Fuks *et al.*, 2007). La coordination prévient tout conflit ou redondance dans l'accomplissement de la tâche (Pimentel *et al.*, 2004), la communication permet aux participants de signaler leurs intentions, favorisant la coordination des actions. Communication, coordination et coopération aident à prendre acte des actions produites par les autres participants et à mesurer la progression du groupe dans la tâche collective.

Dans les deux sections suivantes nous présentons une définition plus précise de ce que l'on peut appeler une coopération et une coordination.

5.1.2 Coopération

Fuks *et al.* (2007) définit la coopération comme la « production jointe de membres d'un groupe dans un espace partagé générant et manipulant des objets de coopération dans l'objectif de compléter une tâche ». Dans cette définition, l'accent est mis sur la notion d'action jointe entre les participants. Une action jointe est définie par

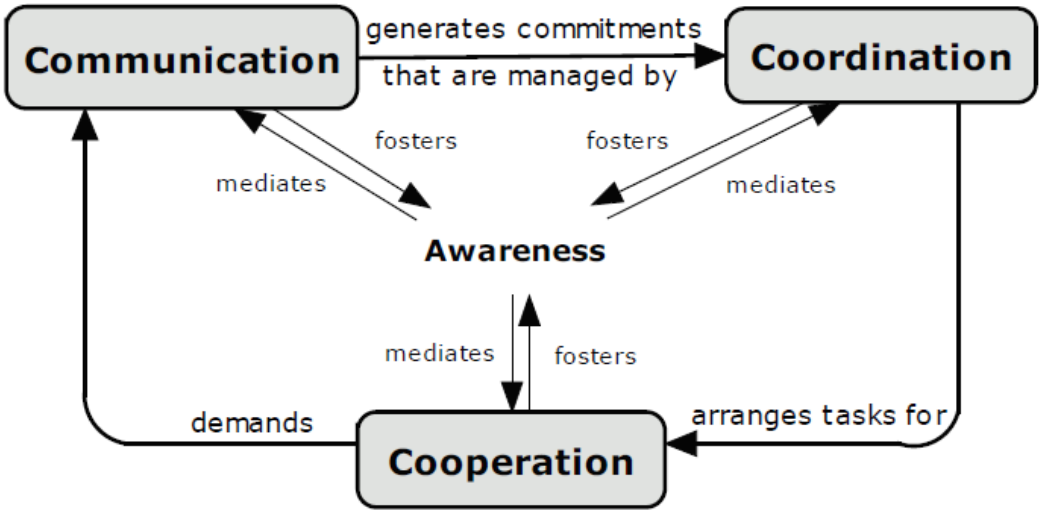


FIGURE 5.1 – Schéma illustrant l'interconnexion entre les trois composants du modèle 3c. Extrait de Fuks *et al.* (2007).

Sebanz *et al.* (2006) comme « une forme d'interaction sociale où deux participants ou plus coordonnent leurs actions pour apporter un changement dans l'environnement ». Ainsi deux participants portant un objet lourd réalisent une action jointe. Tuomela (1993) met en avant une distinction entre la notion de coopération et la notion d'action jointe. Pour cet auteur, l'action jointe nécessite une coordination dans un espace partagé entre plusieurs participants, ce qui n'est pas le cas pour la coopération.

Au-delà de l'accomplissement de tâches procédurales, les actions jointes peuvent prendre plusieurs formes. Directement liée à notre problématique, la production de langage dans une conversation est elle-même considérée par Clark (1996) comme une action jointe. Cette action jointe est coopérative ou non-coopérative selon le type de dialogue. La coopération dans la production de langage a lieu lorsque les auditeurs produisent des *backchannels* pour inciter le locuteur à poursuivre la production de son énoncé (Clark, 1996) ou encore s'observe lors de complétions collaboratives d'énoncés par les auditeurs (Clancy et McCarthy, 2015). Le dialogue laisse entrevoir aussi des actions non-coopératives. Par exemple, deux participants cherchant à prendre le tour dans une situation de conflit de parole, cherchent à prendre ou garder la parole en agissant à l'encontre les intentions de l'autre participant.

5.1.3 Coordination

Dans le modèle 3C, toute action jointe implique une forme de coordination entre les participants (Ellis *et al.*, 1991). Il nous semble donc important de définir ici ce que l'on entend par coordination. Kelso (2009) définit la coordination comme « une forme et un degré d'ordre fonctionnel parmi plusieurs entités et processus dans l'espace et le temps ». Lorsque des entités sont coordonnées, la production de l'action de l'un

est dépendante de la production d'action de l'autre (Kelso, 2009). Pour accomplir une tâche donnée, ou remplir une fonction vitale, différentes entités biologiques ont besoin de se coordonner. Cette coordination est essentielle dans le monde du vivant : sans coordination, cellules, organes, et autres êtres vivants ne pourraient exister. De nombreuses activités humaines requièrent aussi à différentes échelles (petit groupe, entreprise, État par exemple) une coordination entre différents participants (Kelso, 2009). La coordination est fonctionnelle car elle est liée à une activité ou un rôle à remplir. Von Holst (1973) définit trois niveaux de coordination des êtres vivants et systèmes biologiques.

1. Une coordination absolue où les individus sont dépendants mutuellement.
2. Une coordination relative où les individus ne sont pas complètement dépendants ni complètement indépendants. Ils sont partagés entre une tendance à agir de manière autonome et une tendance à dépendre des actions des autres parties, ce que Kelso (2009) appelle la méta-stabilité.
3. Un manque complet de coordination entre les participants.

L'existence de ces formes de coordination est rarement permanente. Deux participants humains peuvent fortement être dépendants dans l'accomplissement de leurs actions lorsqu'ils doivent accomplir une tâche donnée puis arrêter leur dépendance mutuelle lorsque les objectifs des participants sont satisfaits.

5.2 Approches à base de représentations

Comprendre les processus de perception et d'action en situation de collaboration nécessite de comprendre la manière dont un agent biologique perçoit son environnement et décide d'agir selon ces informations de perception. Nous présentons ainsi dans la suite de ce chapitre, les approches de psychologie cognitive modélisant les processus cognitifs en jeu lorsqu'un agent doit accomplir une tâche dans un environnement. Étudier le comportement humain nécessite ainsi d'étudier (Warren, 2006) :

— la coordination de l'action, ou comment, à partir d'une décision d'action, le système neuro-musculo-squelettique s'organise pour réaliser un mouvement ;
— le lien entre perception et action, ou comment l'information au sujet de l'environnement et de son propre corps permet de sélectionner des actions appropriées et adaptées aux conditions environnementales.

Nous pouvons distinguer deux grandes catégories de modèles : l'approche par représentations et les approches situées de la cognition.

L'approche par représentations considère que la tâche de perception consiste à élaborer une représentation interne intelligible pour l'agent de l'environnement, sur lequel l'agent se base pour planifier ses actions (Loomis et Beall, 2004).

Tel que défini par Loomis et Beall (2004), la représentation interne de l'environnement est composée d'un ensemble d'unités de perception, les *percepts* provenant en grande partie de l'information enregistrée par les organes sensoriels. Pour développer sa représentation interne l'agent utilise aussi certaines hypothèses sur l'environnement (rigidité des objets qui se déplacent, continuité du flux optique et des surfaces) et des attentes provenant d'interactions passées avec l'environnement. L'agent planifie ses actions sur la base de ces représentations internes. L'action est conceptualisée dans ces approches comme des schèmes moteurs abstraits divisés ensuite en hiérarchies d'actions qui sont concrètement réalisées dans l'environnement (Schmidt, 1975). Néanmoins, les actions réalisées dans un environnement naturel sont soumises à des contraintes physiques nécessitant un mécanisme de rétroaction pour contrôler la réalisation du mouvement dans l'environnement. Ces mécanismes de contrôle ne s'appliquent pas aux membres contrôlés directement mais la « surveillance » de l'exécution de l'action s'effectue par une représentation interne de l'action en cours d'exécution.

Une majorité des architectures d'agent s'est inspirée des principes de l'approche par représentation. Si l'on considère un système de dialogue, celui-ci n'interprète pas directement les mots prononcés par un utilisateur mais fait correspondre à ces mots des représentations sémantiques que l'agent peut manipuler et sur lesquelles il peut raisonner. Il formule aussi symboliquement la phrase à prononcer qui est ensuite transcrite dans la langue de l'utilisateur. Si l'on considère SAIBA ou *Virtual Human Toolkit* les mêmes principes sont mis en avant : à partir des actions produites par l'utilisateur, une intention communicative est déduite, traitée par l'agent, qui formule en retour une intention communicative en réponse, cette intention étant transformée en actions multimodales réalisées concrètement par l'agent. Les architectures d'agent SOAR (Laird *et al.*, 1987) et ACT-R (Anderson, 1983) sont des architectures générales pour la gestion du comportement d'un agent dans un environnement s'inscrivant dans une approche par représentations. Ces deux architectures se basent sur la « Théorie Computationnelle de l'Esprit » (Pylyshyn, 1981) considérant que la cognition humaine peut être modélisée de la même manière qu'un ordinateur exécute un programme. D'autres approches comme l'approche *Belief, Desire, Intentions* ou *BDI* de Rao et Georgeff (1991) suit aussi les principes de l'approche par représentations.

5.3 Approches situées de la cognition

Les approches dites situées de la cognition offrent une alternative à l'approche par représentations en rejetant l'idée de symboles, représentations, procédures et plan d'actions dans la perception et le contrôle de l'action.

5.3.1 L'approche écologique du comportement

L'approche écologique du comportement est apparue grâce aux travaux de Gibson (1979). Gibson s'oppose aux approches par représentations en reprochant à leurs auteurs le manque d'étude du comportement humain en environnement réel, où le participant perçoit et agit afin d'accomplir une tâche réelle. La perception est souvent étudiée dans des contextes d'expérimentations en laboratoire où le participant est immobile, contrairement à la perception en environnement réel où le participant est au contraire actif, et recherche activement les informations nécessaires à la réalisation de son comportement (Gibson, 1979). Puisque l'agent est constamment en mouvement, l'information provenant de l'environnement est dynamique, évolue en même temps que l'agent agit, ce qui modifie continuellement l'information perçue par l'agent. Pour replacer la perception et l'action dans leur contexte environnemental, l'approche écologique propose une approche « incarnée » et « située » de la cognition (Richardson *et al.*, 2008). Dans cette approche, l'agent possède un corps et se situe dans un environnement. Il subit un certain nombre de contraintes : physiques, liées à l'information provenant de l'environnement, à la mécanique du corps et aux demandes de la tâche à accomplir.

Richardson *et al.* (2008) mettent en avant six principes de l'approche écologique du comportement.

1. Les informations servant à la prise de décision de l'agent ne sont pas des symboles, mais sont directement les stimuli reçus de l'environnement (Gibson, 1979). Plus précisément, l'information est spécifiée par les invariants dans les stimuli récoltés par les organes sensoriels. Les invariants représentent les stimuli restant constants à mesure que l'agent agit dans son environnement. Nous pouvons citer comme exemple d'invariant le flux optique. Lorsqu'un agent se déplace, le flux optique reçu par l'agent sur sa rétine se déplace, excepté dans le cas où l'agent regarde dans la direction de son mouvement. Dans ce cas là, le flux optique reste constant au centre de la rétine de l'agent. Selon Gibson (1979), un agent biologique exploiterait activement cette caractéristique du flux optique pour se diriger vers une cible. Plus précisément, il lui suffirait de faire en sorte que l'image de la cible reste constante sur sa rétine.

   D'autres exemples, comme les variations dans le niveau de détail des textures des objets ou le mouvement parallaxe (le fait que lorsque l'on se déplace, les objets les plus proches de nous apparaissent venir vers nous beaucoup plus rapidement que les objets lointains) donnent des informations sur la distance de l'agent envers ces objets. À mesure que l'agent interagit avec son environnement, il reconnait au fil de ses interactions des invariants de plus en plus complexes. Ces invariants constituent des invites d'action (*affordances*) que l'agent exploite directement pour agir dans son environnement. Ces *affor-*

dances peuvent être positives, une chaise de par sa forme offrant la possibilité de s'asseoir pour l'agent, ou négative, un prédateur étant une invite à fuir (Gibson, 1979).

2. L'information est spécifique à l'environnement, à l'agent et à la tâche. Une *affordance* constitue ainsi une opportunité d'action seulement pour certains agents, et ne seront donc perçus que par ces mêmes agents.
3. Le contrôle du comportement n'est pas centralisé, l'agent ne planifie pas à l'avance ses actions sur la base de représentations internes de l'environnement. Au contraire, l'information guide directement le contrôle de l'action de l'agent. En ce sens, l'information constitue une source de contrainte au même titre que les contraintes physiques. La tâche de l'agent est d'exploiter ces contraintes pour simplifier le contrôle de son action.
4. Les organes sensoriels ne récoltent pas passivement l'information, mais « s'engagent » dans la détection de l'information. Les yeux ne sont, par exemple, pas fixes, mais des organes que l'agent déplace pour rechercher l'information nécessaire au contrôle de ses actions.
5. Il existe une dépendance mutuelle continue entre perception et action. L'action est directement guidée par l'information provenant de l'environnement qui de son côté modifie la configuration des organes sensoriels modifiant la perception et l'information récoltée. Cette dépendance entre perception et action constitue le cycle perception-action.
6. Le comportement est émergent et auto-organisé. Le contrôle du comportement est distribué entre l'agent et l'environnement et ne provient d'aucune structure identifiable dans l'environnement ni dans l'agent. Le système agent-environnement est en ce sens un système complexe, dont l'interaction entre agent et environnement fait émerger le comportement, entité ayant sa propre évolution indépendamment de l'agent ou de l'environnement seul et contraignant les actions de l'agent.

Ce n'est que récemment que l'intelligence artificielle a proposé des architectures appliquant ces principes à la conception d'entités artificielles (Pfeifer et Pitti, 2012). La motivation à sortir du paradigme par représentations provient de la difficulté de cette approche à modéliser certaines actions humaines. L'approche par représentations excelle dans la modélisation de tâches formelles nécessitant peu d'interactions avec l'environnement. Mais lorsque l'on souhaite modéliser des activités qui paraissent simples comme la locomotion humaine, les approches par représentations se heurtent à une explosion combinatoire des solutions. Le corps humain est en effet composé d'un grand nombre de degrés de liberté offertes par les différentes articulations et muscles composant le corps (Pfeifer et Pitti, 2012). Si l'on suit l'approche traditionnelle par représentations, nous nous retrouvons à chercher une solution dans des espaces très grands pour trouver des solutions efficaces au contrôle de la

locomotion. Nous sommes de plus constamment en interaction avec l'environnement subissant ses forces physiques, parfois imprévues. Le problème du contrôle de la locomotion devient alors difficilement abordable par l'approche par représentations (Pfeifer et Pitti, 2012). Au contraire, si l'on s'intéresse à la manière dont le corps est influencé par les forces physiques, on constate que l'élasticité des tendons, la composition des muscles et la structure du squelette humain impose un couplage physique entre certains membres, diminuant les configurations possibles pour réaliser un mouvement : c'est le principe même de synergie qui s'applique ici (Pfeifer et Pitti, 2012).

Une des approches les plus célèbres se détachant de la vision par représentations en intelligence artificielle est celle de l'Intelligence Artificielle Comportementale de Brooks (1986). Brooks propose une architecture où le contrôle comportement global d'un agent est distribué entre des modules spécialisés. Ces modules sont organisés en parallèle et disposés dans différentes couches de contrôle du comportement définissant différents niveaux de priorité. L'information provenant de l'environnement est directement intégrée dans l'architecture et le contrôle du comportement est géré par plusieurs modules de prise de décision spécialisés qui contrôlent directement les actionneurs de l'agent. La sélection de l'action est gérée entre les modules, des modules de priorité plus élevée pouvant subsumer les décisions prises par des modules de priorités moindres. Selon Brooks (1986), reprenant l'idée du caractère superflu des représentations prôné par l'approche écologique, « le monde est son propre modèle ». L'idée d'IA Comportementale a été utilisée dans la conception de l'architecture Ymir de Thórisson (1999).

Suivant l'approche écologique et le principe d'autopoïèse (Maturana et Varela, 1980), l'énaction propose que le fondement même du couplage entre agents biologiques et environnement provient de la nécessité d'un organisme de préserver sa viabilité en réaction aux perturbations de l'environnement. Une perturbation de l'environnement provoque une modification du fonctionnement d'un organisme qui modifie à la fois sa structure interne et ses actions dans l'environnement. Nous retombons alors sur l'idée de couplage agent-environnement, avancée par le principe écologique de la cognition. L'intelligence artificielle basée sur l'énaction propose d'appliquer ces principes à des entités artificielles (De Loor *et al.*, 2009). La justification de l'application d'un tel principe biologique provient de l'idée que le principe de l'énaction pourrait être appliqué à d'autres formes de système tels que des collectifs d'êtres humains (De Loor *et al.*, 2009). Une idée phare de l'approche de l'énaction est la notion d'ultra-stabilité (De Loor *et al.*, 2015) illustrée sur la figure 5.2. Un système ultra-stable est composé de deux boucles sensorimotrices, une première correspond au couplage sensorimoteur au sens de l'approche écologique de la cognition. La seconde boucle de rétroaction préserve les variables dites « essentielles » (De Loor *et al.*, 2015) de l'agent, liées à la préservation de la viabilité de l'organisme dans son

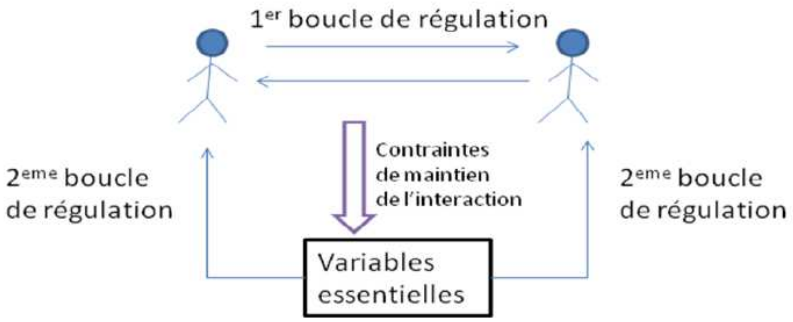


FIGURE 5.2 – Illustration du principe d'ultra-stabilité. Extrait de De Loor *et al.* (2015).

environnement.

5.3.2 Systèmes dynamiques et approche située de la cognition

Parmi les approches situées de la cognition, un intérêt croissant a été porté à l'application des systèmes dynamiques comme outils de modélisation du comportement humain (Warren, 2006; Kelso, 2009). Nous ne rentrerons pas dans le détail du fonctionnement des systèmes dynamiques dans ce chapitre, nous en proposons une introduction en annexe A.

La dynamique de la coordination (Kelso, 2009)

Kelso (2009) formule un cadre d'étude pour la modélisation de la coordination entre éléments d'un système biologique, entre l'humain et son environnement ou dans un système social. L'objectif de son approche, la dynamique de la coordination, est d'expliquer comment ces formes de coordination « apparaissent, s'adaptent, persistent et changent dans les organismes vivants » (Kelso, 2009). Dans le cadre théorique de la dynamique de la coordination, l'accent est mis sur le concept de synergie. Les synergies sont des systèmes biologiques composées d'individus indépendants les uns des autres. Lorsqu'aucune contrainte n'est exercée sur eux, ces individus agissent sans tenir compte des autres, le comportement du système est désordonné. Néanmoins, sous certaines conditions, ces éléments se mettent à agir de manière ordonnée formant une unité cohérente : une synergie. Les possibilités d'action des composantes à l'intérieur du système sont contraintes par le comportement des autres individus du système et par le comportement global de l'unité. Les synergies ont des propriétés essentielles pour garantir la cohérence du comportement de l'unité dans sa globalité.

1. Les composantes sont faiblement couplées entre eux. Elles sont interdépendantes et se coordonnent par le biais de l'information. Contrairement au

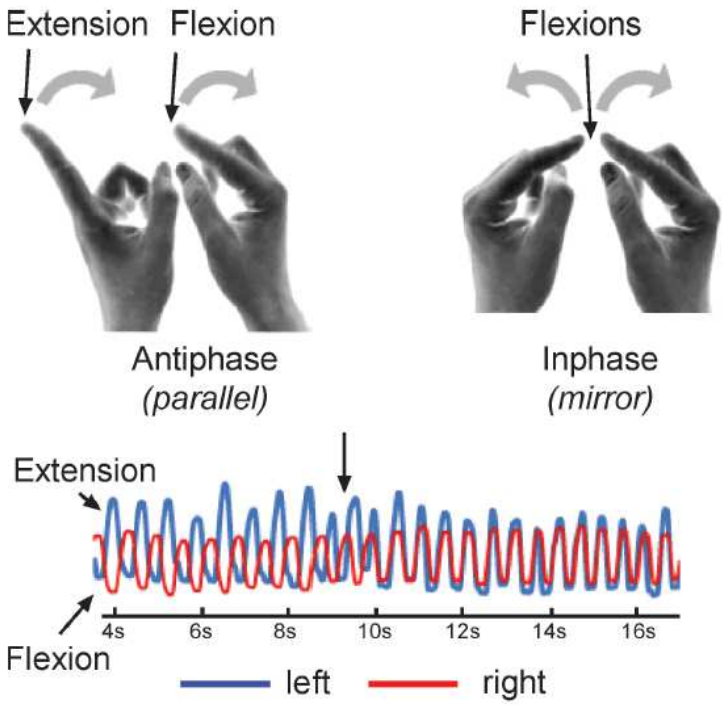


FIGURE 5.3 – Illustration du modèle HKB de Haken *et al.* (1985) modélisant la coordination des mouvements de deux doigts d'un même individu. Extrait de Kelso (2009).

couplage fort, des individus faiblement couplés ont un lien entre eux pouvant varier au cours du temps, et la nature de leur dépendance peut être modifiée. Sous certaines conditions, une perturbation dans le système peut changer la manière dont les éléments interagissent entre eux.

2. Ces systèmes présentent un degré de robustesse élevé. Ils répondent plus précisément au principe d'homéostasie : sous certaines conditions, lorsqu'une perturbation apparait au sein du système cette perturbation est compensée par une modification des actions d'autres éléments non affectés par cette perturbation. Le système garde alors un comportement stable.
3. Ces systèmes rendent compte de l'émergence et de l'auto-organisation du comportement.

Kelso (2009) donne de nombreux exemples de systèmes reposants sur la synergie. Les systèmes cellulaires, ou les systèmes articulatoires et musculaires responsables de la production de parole suivent selon lui le principe de synergie. Afin de rendre compte de ces propriétés, la dynamique de la coordination fait appel au cadre théorique des systèmes dynamiques pour modéliser différents types de systèmes.

Kelso (2009) a appliqué ses travaux à la modélisation de comportements humains. Parmi ces travaux, Haken *et al.* (1985) proposent une expérimentation où les participants oscillent deux de leur doigt de sorte qu'ils soient coordonnés en phase (les doigts sont à la même position au même moment) ou en opposition de phase (lorsqu'un doigt est au maximum de son extension, l'autre est au maximum de sa

flexion). Ils montrent que, peu importe le participant, si le participant coordonne ses doigts en opposition de phase, lorsque la fréquence d'oscillation augmente une transition a lieu vers une coordination en phase. Ce comportement a été modélisé avec succès en suivant les principes de la dynamique de la coordination. Des travaux ultérieurs ont montré qu'une même observation pouvait être effectuée non seulement lors de la synchronisation entre deux doigts d'un même participant mais aussi lorsque le participant doit se coordonner en opposition de phase avec un stimulus sonore (Kelso, 2009). Non seulement il est observé dans cette dernière étude un changement d'un mode de coordination en opposition de phase à un mode de coordination en phase mais le changement de mode de coordination s'effectue lorsque le stimulus est déclenché à la même fréquence que le changement de phase observé par Haken *et al.* (1985).

Kelso *et al.* (2009) ont appliqué avec succès le modèle HKB à une interaction entre un agent artificiel piloté par l'équation HKB et un utilisateur montrant une perspective possible de l'application des principes de la dynamique de la coordination à une interaction utilisateur-agent.

La dynamique comportementale (Warren, 2006)

La dynamique comportementale reprend aussi les principes de l'approche écologique de Gibson (1979) en utilisant des systèmes dynamiques pour modéliser le contrôle des actions de l'agent et son comportement. La dynamique comportementale reprend notamment l'idée de causalité circulaire : l'interaction de l'agent avec son environnement fait émerger le comportement qui en retour contraint les actions de l'agent, ceux-ci permettant de converger vers les objectifs de l'agent. Il est important de noter que l'agent n'est pas passivement contraint par la dynamique comportementale s'établissant au fil de l'interaction, l'agent explore au contraire activement cette dynamique comportementale et l'utilise pour moduler ses lois de contrôle.

La figure 5.4 résume les principes de la dynamique comportementale. À un premier niveau se trouve le cycle de perception-action. À ce niveau, deux systèmes dynamiques interagissent : d'une part, l'agent et d'autre part, l'environnement de l'agent. Plus précisément, l'agent module ses variables d'action par le biais d'une loi de contrôle formulée par l'équation $\dot{a} = \Psi(a, i)$, avec $\dot{a}$ la dérivée des variables d'action, a les actions courantes de l'agent et i les informations récoltées de l'environnement. Les informations récoltées par l'agent correspondent aux invariants de l'environnement (flux optique, texture par exemple). L'environnement a un ensemble de propriétés modifiées par les forces physiques que l'agent exerce et l'évolution de l'état de l'environnement est donnée par l'équation différentielle : $\dot{e} = \Phi(e, F)$, avec $\dot{e}$ la dérivée de l'état de l'environnement, e l'état courant et F les forces physiques exercées par l'agent sur l'environnement. Dans les deux cas, l'information i provenant de

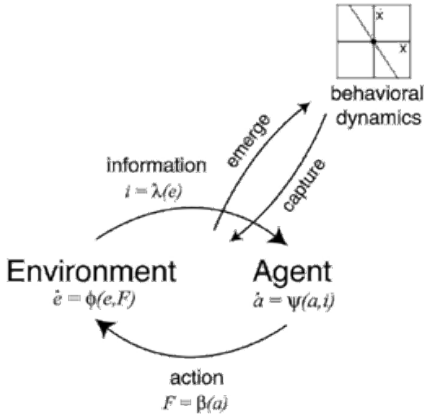


FIGURE 5.4 – Schéma résumant les principes de la dynamique comportementale (Warren, 2006). Extrait de Warren (2006).

l'environnement et les forces physiques F jouent le rôle de variables de contrôle de ces systèmes dynamiques. Le cadre théorique proposé par Warren met en avant le lien direct entre les actions de l'agent et les forces physiques de même qu'un lien entre les propriétés de l'environnement et l'information guidant les actions de l'agent. La loi de contrôle ne formule pas le comportement de l'agent directement, mais la manière dont l'agent va moduler ses variables d'actions en lien avec l'évolution des propriétés de l'environnement. Warren (2006) illustre ce principe en prenant comme exemple le contrôle de la marche. Lorsque l'on marche vers un objectif, nous ne planifions pas en avance l'ensemble de nos mouvements et schémas moteurs pour arriver à notre objectif. Au contraire, lorsque nous commençons à marcher, nous varions l'ensemble des actions de l'agent. Cette mise en mouvement provoque une modification dans l'environnement, perçue par l'agent : le flux optique se déplace sur la rétine, certains objets apparaissent ou disparaissent. Ce changement module en retour les variables d'action provoquant un nouveau changement dans l'environnement et ainsi de suite.

Lorsque l'on parle du comportement dans sa globalité on est à un deuxième niveau d'analyse : la dynamique comportementale. Cette dynamique comportementale est aussi un système dynamique ayant sa propre équation : $\dot{x} = \Omega(x, r)$, avec x l'ensemble des variables définissant le comportement modélisé et r les variables de contrôle du système. Si l'on prend comme exemple le comportement de direction (guidage) vers une cible, les variables de l'ensemble x représentent l'angle entre la direction dans laquelle l'agent se déplace, la direction de la cible et la vitesse à laquelle se déplace l'agent. En comparaison, les variables d'information traitées au niveau du cycle de perception-action sont de l'ordre de la variation de flux optique. r dépend des informations récoltées de l'environnement par l'agent. Si l'on reprend l'exemple d'un agent se dirigeant vers une cible, r varie en fonction de la position de la cible dans le cas d'une cible mouvante. L'espace d'états de la dynamique comportementale définit les buts de l'agent (attracteurs) ainsi que les états à éviter (répulseurs).

Selon la vision de Warren, l'interaction au premier niveau d'interaction donne

naissance à l'espace d'état du comportement (second niveau). L'évolution du comportement dans cet espace d'état rétroagit sur l'agent en modifiant la loi de contrôle et ses variables de contrôle. La dynamique comportementale a été appliquée avec succès à la locomotion vers une cible et à l'évitement d'obstacles (Fajen, 2013), à l'interception d'une cible mouvante (Fajen et Warren, 2007), ou encore à la saisie d'objets à une ou deux mains (Frank *et al.*, 2009). Néanmoins certains types de comportement représentent à ce jour des enjeux de recherche :

- les comportements séquentiels sont des comportements souvent composés de séquences d'actions définies par une série de sous-tâches ou sous-buts ;
- les comportements d'anticipation sont des comportements dont le but de l'agent n'est pas encore présent dans l'environnement, et l'agent ne peut s'appuyer sur de l'information présente dans l'environnement pour guider son action ;
- les comportements prédictifs, liés à des informations de l'environnement non spécifiées directement par les stimuli reçus mais qui correspondent à des associations apprises entre une ou des propriétés « cachées » de l'environnement et propriétés « visibles » de l'environnement. Le caractère cassant d'un objet en verre n'est, par exemple, pas spécifié dans les stimuli de l'environnement, mais l'association verre et cassant est apprise par l'agent ;
- les comportements stratégiques : liés à une connaissance contextuelle de ce qu'il se passe dans l'environnement. Une interaction entre deux agents peut être un comportement stratégique dans le sens où l'agent semble prévoir ce que l'autre va faire en fonction de l'historique de ce que l'agent a réalisé avant.

5.4 Modélisation des processus de coopération et de coordination

Cette section met en avant plusieurs modèles définissant la nature de la coopération entre les participants et des processus cognitifs en jeu lorsque des participants coopèrent.

Théorie des jeux

Une majorité d'approches mettent en avant la modélisation des états mentaux de l'autre et l'utilisation de plans d'action partagés entre les participants suivant une approche par représentation pour comprendre la coopération entre plusieurs participants (Galantucci et Sebanz, 2009). Plusieurs approches (Colman, 2003; Lewis, 1969; Bicchieri, 1989; Binmore, 1987; Caelen, 2003) ont proposé l'abstraction des problématiques liées à la coopération par l'utilisation de la théorie des jeux (von Neumann et Morgenstern, 1947). La théorie des jeux est une théorie mathématique considérant que toute forme d'interaction (sociale, ou autre) peut être modélisée sous

forme de jeux. Son cadre formel provient de la théorie de la décision, une théorie probabiliste dédiée à l'analyse du comportement humain dans le cadre de prises de décisions non-stratégiques et incertaines. Le terme « jeu » doit être compris dans ce cas de figure comme une abstraction mathématique, une idéalisation d'une interaction sociale (Colman, 2003). La théorie des jeux a été étudiée dans d'autres champs disciplinaires notamment l'économie (Rabin, 1991), les sciences politiques (Brams, 1975) et l'impact de la coopération inter-espèces dans la théorie de l'évolution (Axelrod, 1981). Dans cette théorie, les agents impliqués dans le jeu sont rationnels (Colman, 2003) : si un agent a le choix entre plusieurs solutions il choisira toujours la solution maximisant l'*utilité attendue* (von Neumann et Morgenstern, 1947), c'est-à-dire l'alternative maximisant le gain pondéré par la probabilité que cette alternative ait lieu. En ce sens un jeu est modélisé par une matrice de gains, représentant les gains de chaque personne en fonction des choix conjoints de chaque participant. Les participants disposent de plusieurs stratégies possibles et obtiennent des gains plus ou moins grands en fonction de la stratégie choisie et de celle des autres. Pour les auteurs de cette approche (Colman, 2003) les comportements de chaque participant peuvent ainsi être prédits en considérant les choix de chaque participant.

La théorie des jeux est une théorie à la base normative, elle décrit la décision optimale d'agent rationnel mais pas ce qu'un humain choisit en réalité. La contradiction entre l'hypothèse de rationalité et la réalité montre que d'autre facteurs influencent la prise de décisions dans le cadre d'activités coopératives. L'approche prise par la théorie des jeux est individu-centré : elle modélise séparément la prise de décision des deux agents sans prendre en compte le contexte social dans lequel le participant est situé.

La cognition sociale

Des approches plus récentes (Galantucci et Sebanz, 2009) postulent une influence directe du groupe sur la cognition de l'individu qui pourrait expliquer les différences entre les résultats formels observés dans la théorie des jeux et la réalité. Levine et Resnick (1993) mettent en avant cinq manières dont les facteurs sociaux influencent la cognition.

1. La simple présence d'autres personnes dans l'environnement de l'individu influe elle-même sur la cognition, que les autres personnes soient engagées ou non dans une activité coopérative avec l'individu. Cette présence a des effets d'inhibition ou de facilitation sociale, augmentant ou diminuant la performance dans l'accomplissement d'une tâche, provoque une surcharge cognitive dans certains contextes, de l'incertitude et impose des contraintes comportementales.
2. Les rôles et les positions sociales sont sources de contraintes dans le comportement. Le comportement peut être contraint par les attentes qu'on lui

attribue.

3. Les représentations mentales que l'on se fait de l'autre influent de même sur le comportement. Un individu peut modifier son opinion pour se conformer, ou être en opposition, avec un groupe de référence. Nos propres perceptions et évaluations de nous peuvent de même être influencées par une comparaison avec d'autres individus. Nos activités cognitives sont enfin influencées par anticipation d'une interaction avec d'autres individus.
4. L'interaction sociale induit des changements cognitifs. Ces changements cognitifs peuvent être dus à des effets de conformité au groupe social (l'individu en position de minorité tend à adopter les opinions des individus en position de majorité) ou des effets d'innovation (l'individu en position de majorité tend à adopter la position de la minorité). Cette influence n'est pas seulement liée aux opinions de l'individu mais peuvent même influencer jusqu'aux processus perceptifs. Certains auteurs adoptent une position plus radicale en considérant que nos propres capacités fondamentales de penser et nos pensées elle-mêmes sont créées par l'interaction sociale (Vygotsky, 1978).
5. La cognition est partagée dans un groupe et les participants doivent maintenir une compréhension mutuelle de l'environnement et des activités du groupe. Les participants ne sont donc pas seulement influencés malgré eux par le groupe social mais la coordination des perceptions, décisions et actions d'un groupe est un pré-requis nécessaire au succès de la collaboration entre individus. Lorsque les participants perçoivent un stimuli ambigu, ils doivent coopérer pour que les jugements des membres individuels convergent vers un jugement partagé du stimuli.

Ces cinq principes montrent qu'une action coopérative est avant tout distribuée selon les participants. Ceux-ci partagent leur perception et l'action est distribuée entre les participants, chacun occupant un rôle particulier dans la coopération. On ne peut étudier le comportement des individus dans un collectif sans prendre en compte l'influence du collectif sur l'individu. Selon cette perspective de cognition sociale, deux approches s'affrontent. La première, suivant une vision par représentations de la perception et de l'action, considère que le succès de l'action jointe entre deux participants repose sur le partage de représentations mentales entre les participants (Sebanz *et al.*, 2006). Sebanz *et al.* (2006) considèrent ainsi que la coordination des actions dans le cadre d'une action jointe provient de la capacité du participant à observer et déduire les objectifs de l'autre. C'est à ce prix qu'un individu est capable de sélectionner l'action complémentaire à réaliser pour réaliser l'action jointe.

Marsh *et al.* (2006) modélisent la coopération selon les principes de l'approche écologique, en considérant que deux agents sont couplés de la même manière qu'un agent peut être couplé avec son environnement. Plutôt que de proposer des mécanismes d'inférences sur les objectifs d'un partenaire potentiel (Sebanz *et al.*, 2006),

ce partenaire représente directement une affordance de coopération pour l'individu qui perçoit directement l'opportunité d'agir avec ce dernier. L'action jointe est vue selon cette approche selon le principe de mutualité : de la même manière qu'un agent en interaction avec son environnement (Warren, 2006), l'individu perçoit des invariants dans les actions entreprises par l'autre participant. Ces invariants influent directement sur les actions de l'agent qui en retour module ses propres variables d'action.

5.5 Vers un modèle de la prise de parole

Le manque de preuves en faveur de l'approche par projection ou par réaction aux signaux dans les modèles existants de psychologie sociale nous a poussé à nous intéresser de plus près aux modèles de psychologie cognitive explorant les processus en jeu dans une action jointe. De l'exploration de ces modèles, nous pouvons constater en premier lieu que les modèles de gestion du tour de parole suivent en grande majorité l'approche par représentation. Les rares modèles sortant de cette approche (Wilson et Wilson, 2005; Ikegami et Iizuka, 2007) sont soit critiquables dans leurs postulats (Wilson et Wilson, 2005), soit s'intéressent à des mécanismes de gestion de tour de manière générale sans adresser spécifiquement la manière dont les participants engagés dans une conversation s'échangent leur tour (Ikegami et Iizuka, 2007). Si l'on se penche de plus près sur le déroulement de la conversation on constate plusieurs choses. Les participants agissent et réagissent en continu aux actions de l'autre. L'auditeur fournit des *feedbacks* au locuteur sur sa compréhension de l'énoncé ou l'interrompt si nécessaire, le locuteur en retour ne planifie pas sa phrase mais la construit à mesure que l'autre participant lui fournit des indices sur sa compréhension (Clark, 1996). Il en résulte des situations où le locuteur courant produit des *fillers* pour réfléchir à ce qu'il pourrait dire tout en gardant le tour, ou des situations où le locuteur, en conflit de parole avec l'auditeur, répète la dernière syllabe qu'il prononçait jusqu'à ce que l'auditeur lui laisse le tour ou qu'il abandonne le tour. Les participants s'ajustent donc en temps réel au comportement de l'autre. Cela implique que les comportements des participants ne sont pas prévisibles à l'avance mais sont nécessairement émergents de l'interaction entre les participants. Modéliser la gestion des échanges de paroles comme un processus continu et émergent fournirait une alternative viable aux modèles existants pour expliquer comment les participants coordonnent finement leurs échanges de tour. Un tel modèle expliquerait aussi la capacité des participants à résoudre rapidement les conflits de parole ayant lieu dans la conversation, en considérant que ces derniers ne projettent pas la fin de tour, ni ne réagissent à des signaux événementiels mais perçoivent de manière continue les signaux de l'autre, et plutôt que de prendre des décisions comportementales symboliques, moduleraient directement leurs actions

motrices en réponse. Une approche située de la coordination des échanges de parole semble donc plus appropriée. Néanmoins, la prise de tour apporte un certain nombre de problématiques qui doivent être résolues pour la modélisation de manière située de la prise de tour. L'agent n'a ainsi pas accès directement aux objectifs de son interlocuteur, il le perçoit indirectement par les signaux de son partenaire. On se situe donc dans le cadre d'un comportement prédictif selon les termes de Warren (2006), l'agent doit percevoir une information qui lui est cachée. Pour percevoir la nature du comportement de son interlocuteur l'agent associe indirectement les signaux avec le comportement de l'agent. Cette association est réalisée dès le début de la réalisation des signaux de fin de tour de son interlocuteur. L'agent a donc à ce stade une information encore incertaine sur le comportement de son partenaire. Cette problématique ne peut être résolue par l'emploi de modèles complets actuels de perception et d'action suivant l'approche incarnée (Kelso, 2009; Warren, 2006; De Loor *et al.*, 2009). Néanmoins de récents travaux se sont focalisés sur la modélisation de la prise de décision perceptive de manière incarnée, c'est-à-dire la manière dont les participants prennent des décisions sur la nature d'une information incertaine et ambigüe. Nous présentons ces travaux dans le paragraphe suivant.

Prise de décision perceptuelle face à une information ambigüe

Lorsque qu'un agent détermine si un partenaire prend le tour ou non, ou s'il est en train de laisser le tour ou non, il prend une décision sur la nature de l'information qu'il a devant lui entre deux alternatives, en temps limité et sous conditions d'incertitude concernant la nature du comportement son partenaire. Cette forme de prise de décision constitue une tâche de choix forcé à deux alternatives (*two alternative forced choice* ou *TAFC* en anglais) (Bogacz *et al.*, 2006). Sous ces conditions d'applications, des modèles formels rendant compte de la prise de décision ont été développés. Ces modèles émettent trois hypothèses sur la prise de décision (Bogacz *et al.*, 2006) :

1. le processus de décision s'effectue en accumulant de manière continue des indices provenant de l'environnement ;
2. celui-ci est soumis à des fluctuations aléatoires ;
3. une décision est prise lorsque l'agent a récolté suffisamment d'indices favorisant l'une ou l'autre des alternatives.

Les modèles de *TAFC* sont tous bâtis de la manière suivante. La quantité d'indices favorisant chaque alternative varie au cours du temps. L'agent évalue la différence dans les quantités d'indices récoltées favorisant l'une ou l'autre des alternatives. Lorsque la différence dans la quantité d'indices atteint une valeur suffisante, une prise de décision définitive en faveur d'une alternative par rapport à l'autre est prise. L'alternative ayant récolté le plus d'indices est alors prise. Il existe plusieurs modèles de TAFC néanmoins sous certaines conditions d'optimalité, Bogacz *et al.*

(2006) montrent que ces modèles peuvent se réduire au modèle de de dérive-diffusion ou DDM formalisé par Ratcliff (1978). Nous détaillons le fonctionnement du DDM en annexe B.

Le *DDM* a été appliqué à la modélisation d'un certain nombre d'activités humaines dont la discrimination entre deux types de stimuli visuels (Ratcliff et Rouder, 2000) ou la nature du mouvement d'objets dans un environnement (Ratcliff et McKoon, 2008). Le modèle rend compte à la fois de données expérimentales en psychologie cognitive et en neurophysiologie (Ratcliff et McKoon, 2008). Le *DDM* ne se préoccupe que de la prise de décision de l'agent, sans s'interroger sur les liens entre prise de décision et action.

Lepora et Pezzulo (2015) proposent un état de l'art des différentes approches modélisant la prise de décision par accumulation et le lien avec l'action de l'agent. Les modèles séquentiels traitent la décision et l'action comme des entités séparées, avec l'action réalisée après que l'agent ait franchi un seuil de décision marquant sa décision définitive vis-à-vis d'une alternative. Les modèles parallèles proposent une action réalisée en même temps que la décision. À mesure que la décision évolue, l'action est continuellement révisée. Parmi cette catégorie de modèle, certains modèles considèrent que l'action de l'agent provient de deux commandes motrices simultanées, correspondant aux actions prises selon les deux alternatives et ensuite fusionnées pour contrôler l'action. L'action est modifiée en changeant les poids des commandes motrices selon l'évolution de la décision. Une troisième catégorie de modèles proposent non seulement une génération continue de l'action en parallèle de la décision mais aussi une rétro-action directe de l'action sur le processus de décision. Lepora et Pezzulo (2015) motivent ce choix en tenant compte du fait que l'action contraint autant la perception que l'inverse, en accord avec les principes de la cognition incarnée. La mise en action de l'agent a un certain coût temporel et physique, que l'agent doit prendre en compte dans sa prise de décision. Si le changement d'action a un coût élevé, parce que l'agent ne dispose pas d'assez de temps ou changer d'action est trop coûteux physiquement l'agent aura tendance à rester sur la même décision. Les auteurs proposent une comparaison de ces modèles pour une tâche où des participants doivent choisir entre deux stimuli visuels. Pour observer la dynamique de la prise de décision il est demandé aux participants de sélectionner à la souris, entre les deux stimuli, celui qui correspond à un mot. Les trajectoires des pointeurs de souris sont enregistrées. La figure 5.5 montre un exemple d'une retranscription des trajectoires des curseurs de la souris par les expérimentateurs. Par comparaison entre des simulations des différents modèles, les auteurs montrent que le modèle « incarné » donne la trajectoire moyenne la plus proche des données expérimentales, ne se distinguant que peu significativement du modèle où l'action est modulée en continu selon la décision du participant.

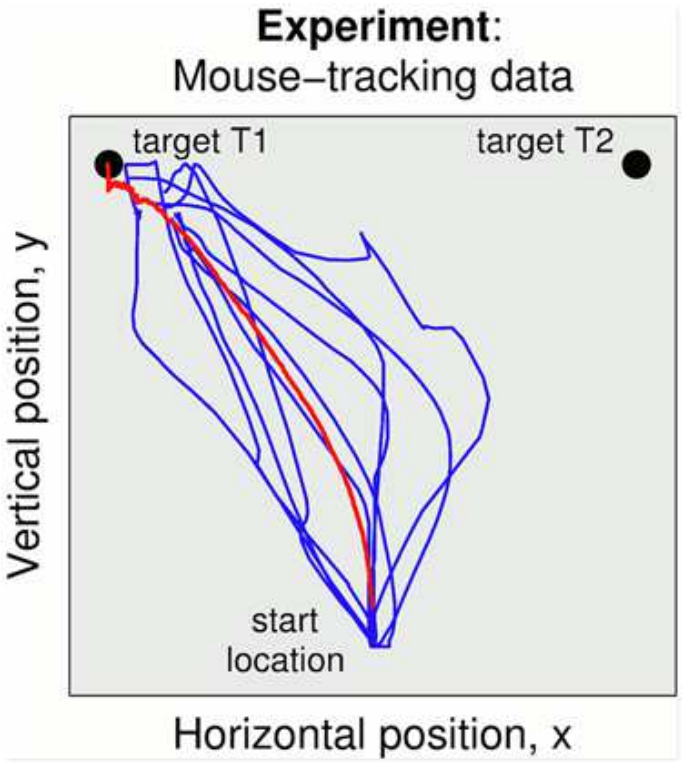


FIGURE 5.5 – Exemple de trajectoires du pointeur d'une souris dans une tâche de discrimination d'un stimuli visuel. Extrait de (Lepora et Pezzulo, 2015).

Une dynamique comportementale de la gestion de tour

La modélisation par systèmes dynamiques de la perception proposé par Warren (2006) et Kelso (2009) est intéressante pour la modélisation de la gestion du tour de parole entre un agent artificiel et l'utilisateur. Ils rendent compte des principes d'une approche située de la cognition en fournissant le cadre théorique nécessaire pour concevoir les lois de contrôle d'un agent en interaction avec l'utilisateur.

Kelso (2009) a d'ailleurs proposé un partenaire virtuel interactif représenté par une main, cordonnant le mouvement de son index avec l'utilisateur. De son côté, Rio *et al.* (2014) ont modélisé la dynamique comportementale de piétons marchant en groupe et suivant un leader, rendant compte de la coordination de la vitesse et de la direction des individus. Warren (2006) propose une formulation détaillée de la manière dont l'information agit en tant que variable de contrôle des actions de l'agent, tandis que Kelso (2009) ne propose pas de description de son cadre théorique à ce niveau de détail. Néanmoins, dans une conversation, les participants perçoivent et agissent sous contrainte de temps, et nous émettons l'hypothèse que les participants agissent en continu, même lorsqu'ils n'ont peu de signaux leur indiquant la fin ou la prise de tour d'un participant. Le cadre de la prise de décision perceptuelle (Lepora et Pezzulo, 2015) et plus particulièrement l'emploi d'un modèle parallèle nous semble approprié pour la perception du comportement de l'interlocuteur. L'agent n'attend pas la fin de tour ou le début de tour d'un participant pour commencer à produire ses propres actions. Lorsque son niveau d'incertitude est grand, ses actions

suivent deux tendances inverses (prendre le tour ou non par exemple) résultant en une modulation d'action hésitante entre ces alternatives, mais à mesure que le niveau de certitude augmente, l'agent module de plus en plus ses actions en faveur d'une alternative par rapport à l'autre.

Un modèle parallèle de la prise de décision et de l'action entre dans le cadre de la dynamique comportementale. La dynamique comportementale telle que formulée par Warren (2006) ne rend pas compte d'une rétro-action directe de l'action sur la perception, mais l'influence de l'action sur la perception est indirecte, médiatisée à travers le changement que produisent les actions sur l'environnement de l'agent. Le modèle parallèle de la prise de décision telle que présentée par Lepora et Pezzulo (2015) semble donc plus approprié qu'un modèle de choix incarné.

Deuxième partie

Modèle continu et émergent de coordination du tour de parole

Chapitre 6

Modèle théorique

Nous présentons maintenant la conception de notre modèle de gestion du tour de parole. Nous rappelons que nos objectifs sont de concevoir un modèle de gestion du tour de parole pour un agent conversationnel engagé dans une interaction dialogique dyadique avec un utilisateur. Chaque participant, utilisateur ou agent, a un certain degré d'autonomie vis-à-vis de l'interaction. Leur but à un moment donné de la conversation peut être de parler ou de laisser le tour. Ces buts sont plus ou moins contradictoires avec ceux de leur partenaire, ce qui génère différentes situations communicatives que l'on peut associer à la gestion du tour de parole, comme les transitions fluides, les moments de silence et les conflits (Heldner et Edlund, 2010). Nous souhaitons reproduire ces situations par l'interaction entre deux agents et entre un agent et un utilisateur.

Le défi est de proposer un modèle plausible rendant compte de la manière dont les participants humains s'échangent la parole, et qui soit une alternative aux approches par projection et par réaction aux signaux. Pour la conception de ce modèle, nous adoptons une approche différente s'inspirant de l'approche de la dynamique comportementale et du paradigme des tâches de choix forcé à deux alternatives pour proposer un modèle entièrement continu et émergent de la gestion du tour de parole. Dans ce chapitre, nous nous intéressons au modèle théorique de gestion du tour de parole, sans nous intéresser à son implémentation dans un système réel de dialogue.

Ce chapitre est divisé en deux sections. Nous proposons dans une première partie des hypothèses sur la manière dont les participants gèrent leur tour dans une conversation, comblant le manque entre des modèles descriptifs du tour de parole et un modèle décrivant les processus de perception et d'action engagés dans la coordination du tour de parole. Nous décrivons ensuite en détail notre modèle.

6.1 Hypothèses d'implémentation du modèle

6.1.1 Dépendance au contexte

La gestion du tour de parole est généralement vue comme un « mécanisme » (Sacks *et al.*, 1974) indépendant du sujet et du type de la contribution verbale (informer ou signifier son désaccord par exemple) ainsi que du contexte social et environnemental dans lequel il a lieu. Pour Sacks *et al.* (1974), peu importe le contexte, les mêmes règles énoncées dans le chapitre 4.2.2 s'appliquent et la nature du processus de perception est la même. Nous rejoignons cette vision en considérant que la manière dont les participants interprètent et produisent des signaux pour se coordonner est sans doute propre à la gestion du tour de parole et indépendant d'autres facteurs contextuels. Néanmoins, et suivant les arguments d'O'Connell *et al.* (1990) et de Clark (1996), les buts de l'agent dans le cadre de la coordination de la parole (prendre le tour, laisser le tour, garder le tour) sont, eux, liés au contexte de la conversation, notamment au contenu verbal échangé entre les participants. De même, le type de conversation et le degré d'affinité des participants dans la conversation ont clairement un rôle dans la distribution des échanges de parole (O'Connell *et al.*, 1990). Nous ne nous intéressons pas aux nombreux facteurs pouvant influencer les échanges de tour, nous considérons que ces facteurs influencent principalement les buts initiaux de l'agent à prendre, laisser ou garder le tour. Ces buts influent sur la variation des actions de l'agent par l'intermédiaire d'une variable continue que nous appelons ici motivation allant d'une forte motivation à ne pas parler à une forte motivation à parler. Nous formulons donc l'hypothèse suivante :

Hypothèse 1. *L'occurrence et la durée d'une transition, ou d'un conflit sont dépendants des motivations à parler des deux participants.*

Nous nous intéresserons, dans la suite, à la manière dont cette motivation modifie la manière dont les participants coordonnent leur parole mais nous ne modéliserons pas la manière dont la motivation à parler varie selon l'état du dialogue ou l'état cognitif du participant.

6.1.2 Caractère bidirectionnel des échanges de parole

Un consensus existe quant au caractère distribué des échanges de parole entre les participants dans les modèles de gestion du tour de parole. Sacks *et al.* (1974) voient la gestion des tours comme un processus laissé à l'administration des participants et contrôlé par l'interaction. Néanmoins dans l'approche de Sacks *et al.* (1974), l'occurrence d'une transition de tour est avant tout à l'initiative du locuteur courant : ce dernier fournit dans son tour des *TCU* permettant aux auditeurs de projeter un moment approprié pour prendre le tour. À l'inverse, Duncan (1972) considère qu'un changement de tour peut autant être à l'initiative du locuteur que de l'auditeur, et

ne mentionne pas de moments particuliers pour initier une prise de tour de la part d'un auditeur. Nous suivons ce postulat et formulons l'hypothèse suivante :

Hypothèse 2. *L'occurrence d'un changement de tour peut autant être à l'initiative de l'auditeur que du locuteur.*

6.1.3 Multimodalité

Lorsque les participants sont engagés dans une conversation, ils produisent un ensemble de signaux informant leur partenaire de leur prise ou de leur abandon de parole. Ces signaux sont coordonnés temporellement, mettant en avant l'existence d'une synergie dans la production multimodale de ces signaux émergeant de l'interaction entre les agents et contraint par les motivations à parler ou non des participants. De même, lorsqu'un participant est engagé dans une conversation, il surveille l'opportunité d'une prise de parole par l'interprétation multimodale d'un ensemble de signaux de fins de tour (Gravano et Hirschberg, 2011) pour l'auditeur, ou est à l'affut de signaux de prise de tour pour le locuteur. Nous formulons cela par les hypothèses suivantes.

Hypothèse 3. *Un participant perçoit les motivations de parler de l'autre en interprétant l'ensemble des signaux produits par l'autre.*

Hypothèse 4. *Un participant coordonne sa production d'action de sorte de signaler sa motivation de parler ou non.*

6.1.4 Accumulation continue de signaux

Dans la section 5.5, nous avons émis l'hypothèse que les participants interprétaient non pas un ensemble de signaux événementiels comme proposé par Sacks *et al.* (1974) ou Duncan (1972) mais plutôt des variations continues de grandeurs prosodiques ou non-vocal. Il n'est pas toujours aisé de percevoir le comportement d'un participant vis-à-vis de la prise ou de l'abandon de parole et il est probable que la perception du comportement de l'autre participant ne soit pas binaire. Le participant serait plus ou moins certain de la nature du comportement du partenaire. Aussi, de notre point de vue, la prise de décision de l'agent pourrait suivre les principes d'accumulation continue mis en avant par les modèles de *TAFC* (*Two Alternative Forced Choice Tasks*) (Bogacz *et al.*, 2006). L'agent varierait son degré de certitude concernant le comportement de son partenaire en accumulant en continu des indices verbaux ou non-verbaux de son partenaire. Lorsque l'agent accumule suffisamment d'indices favorisant une alternative, c'est-à-dire lorsqu'il atteint un certain niveau de certitude, il prend une décision définitive sur la nature du comportement de son partenaire. Une fois cette décision effectuée, l'agent recommence un nouveau processus de perception, indépendant du processus précédent. Ce principe est résumé par l'hypothèse suivante :

Hypothèse 5. *La perception du comportement de prise ou d'abandon du tour se fait de manière continue par accumulation d'indices provenant de toutes les actions produites par l'autre.*

6.1.5 Couplage sensorimoteur auditeur-locuteur

La dynamique comportementale de Warren (2006) et la notion de mutualité introduite par Marsh *et al.* (2006) mettent en avant l'idée que deux participants dans une action jointe sont couplés. Dans le cadre de la gestion du tour de parole, nous implémentons ces idées en considérant que la variation même des actions de prise ou de fin de tour d'un participant est directement influencée par la variation des actions de l'autre participant. La variation des actions du participant provient d'une interaction complexe entre l'évolution de sa motivation à parler et la dynamique des actions de l'autre participant. Ce que nous appelons « comportement », c'est-à-dire l'évolution au cours du temps des actions du participant est un processus émergent de l'interaction entre l'agent et son partenaire.

Dans l'exemple d'un auditeur ayant envie de parler et de s'engager dans un conflit avec le locuteur courant, c'est à la fois la force de la motivation de l'individu et la force des signaux du locuteur indiquant sa volonté de garder le tour qui va déterminer l'issue du conflit. Si le locuteur produit une faible augmentation de volume sonore et de hauteur de voix, un auditeur avec une motivation forte va prendre le tour au locuteur courant, à l'inverse, un locuteur produisant de fortes variations de signaux prendra le dessus sur un auditeur ayant une motivation faible à parler. De la même manière, la durée de transition sera plus ou moins courte selon les motivations conjointes des participants.

Un cas extrême concerne la motivation de ne pas parler de l'auditeur lorsque le locuteur cherche à lui donner le tour. Malgré une motivation de ne pas parler, la force de la motivation à laisser la parole du locuteur courant pourra contraindre l'auditeur courant à parler en produisant éventuellement un *filler* s'il n'a rien à dire pour le moment. Cette situation est justifiée par certaines observations sur les interactions humaines (Torreira *et al.*, 2015) montrant que la force des signaux du locuteur courant peut pousser un participant à prendre la parole même s'il n'avait pas prévu de la prendre au départ. L'auditeur a une forme de pression à parler qui le pousse à indiquer au locuteur sa prise de tour avant qu'il n'ait pu planifier ce qu'il va dire. Dans ce cas de figure, l'agent produit un *filler* ou une inspiration. À l'inverse une motivation très forte de ne pas parler de l'auditeur pourrait pousser le locuteur à reprendre la parole. Nous résumons ce principe de la manière suivante :

Hypothèse 6. *Les comportements des participants au cours de l'interaction sont émergents de leur interaction mutuelle.*

6.1.6 Incertitude et modulation des actions

Nous avons mentionné deux principes dans les sections 6.1.4 et 6.1.5 : celui d'accumulation de signaux et celui de couplage sensorimoteur entre les participants. Le couplage sensorimoteur (Warren, 2006) implique une modulation directe et continue des variables d'action de l'agent en fonction de l'information provenant de l'environnement. Comme application immédiate de ces deux principes, nous postulons que le niveau de certitude sur le comportement du partenaire module directement la variation des actions de l'agent. L'influence de la certitude sur la variation d'action permettrait d'expliquer la rapidité des transitions de tour sans faire référence à des processus de prédiction par les participants ou de réaction aux signaux. Les participants perçoivent les signaux de fin de tour dès le début de leur occurrence, pouvant apparaitre jusqu'à une seconde avant la fin de tour du locuteur selon Gravano et Hirschberg (2011). Ils varient alors leur niveau de certitude sur la fin de tour du locuteur au cours du temps selon les variations des signaux de leur partenaire qui pourrait atteindre son maximum à quelques millisecondes de la fin de tour de ce dernier (Grosjean et Hirt, 1996). L'agent n'a ainsi pas besoin d'attendre la fin des signaux de fin de tour pour commencer à les interpréter et y réagir. Cette variation des actions se fait en accord avec la motivation à parler de l'agent : dans le cas d'une réaction à une prise de tour, si le locuteur courant est motivé à garder le tour, il cherchera à empêcher l'auditeur courant de prendre le tour. À l'inverse, s'il est motivé à laisser le tour, il laissera l'auditeur courant prendre le tour. Nous résumons ainsi :

Hypothèse 7. *Les actions de l'agent sont directement modulées par, d'une part la perception du comportement du partenaire et d'autre part par la propre motivation de l'agent à parler.*

6.2 Présentation des composantes du modèle

Nous détaillons dans cette section les différentes composantes du modèle. Nous nous intéressons ici uniquement à la manière dont les actions de l'agent sont modulées en continu selon les principes de la dynamique comportementale et comment le comportement de prise de parole ou d'abandon de tour est directement perçue en agrégeant les signaux produits par le partenaire de l'agent. Nous ne présupposons pour l'instant pas de la forme des signaux échangés. Nous nous intéresserons à cette problématique dans la partie III de ce manuscrit. Nous commençons par présenter le modèle et la manière dont perception et contrôle de l'action sont couplés selon les hypothèses de la section 6.1. Nous présentons ensuite en détail chaque composante de l'architecture : dans un premier temps la perception du comportement de parler de l'agent et dans un second temps la modulation des actions de l'agent.

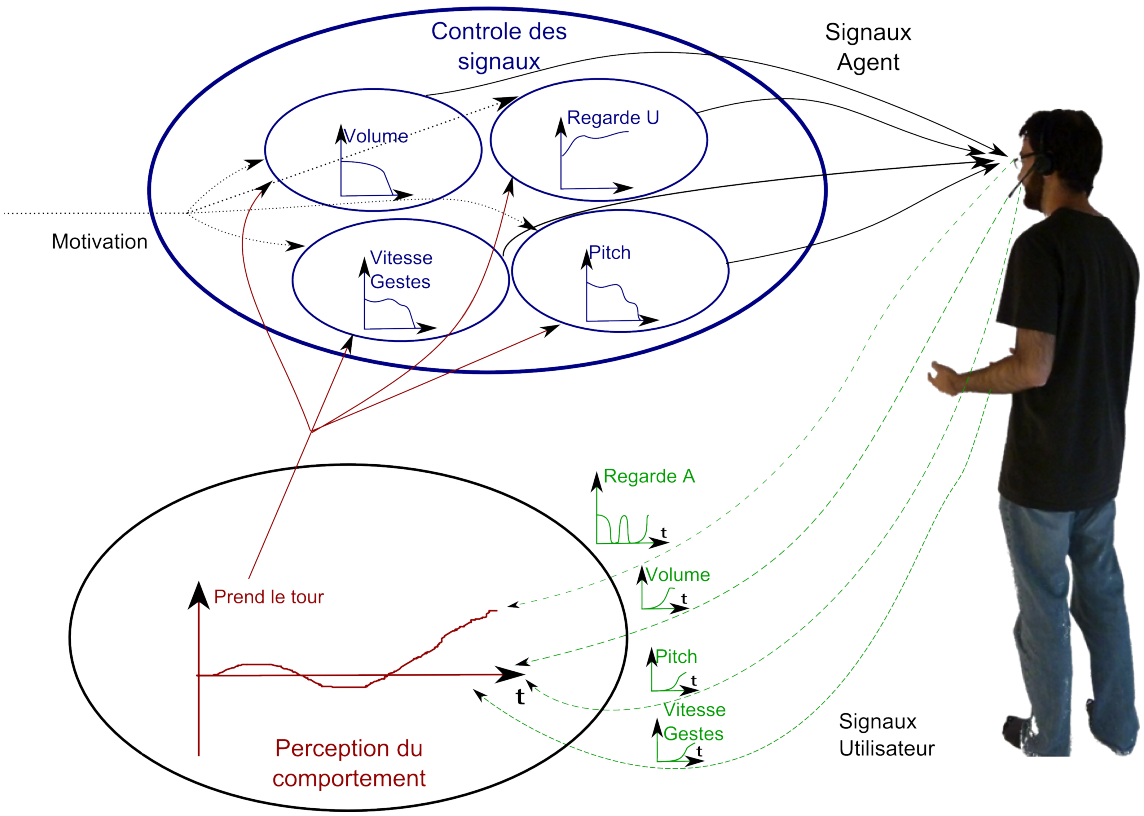


FIGURE 6.1 – Schéma illustrant le principe du modèle théorique.

6.2.1 Modèle théorique

Le modèle général est présenté sur la figure 6.1. Le modèle est composé de deux parties. La partie de perception du comportement de l'agent agrège les signaux provenant du partenaire et met à jour en continu la variable de perception du comportement du partenaire de l'agent.

Le contrôle des signaux de l'agent est guidé conjointement par la motivation propre de parler (ou non) de l'agent et par son niveau de certitude sur le comportement de l'autre agent. L'agent a deux ensembles d'équations de perception et de contrôle des actions. Lorsque son partenaire est locuteur, l'agent perçoit le comportement de son partenaire envers l'abandon de parole et a une motivation à prendre ou non la parole, il contrôle alors ses actions selon cette motivation. Lorsque son partenaire est auditeur, il a un niveau de certitude sur la prise de parole ou non de son partenaire et a une motivation à laisser ou non la parole. Il module ainsi ses actions dans l'objectif de laisser ou garder la parole.

6.2.2 Perception du comportement du partenaire

Nous reprenons les principes du modèle de dérive-diffusion pour l'élaboration du module de perception du comportement du partenaire. Nous reprenons l'équation

du DDM B.1 présentée en annexe B, page 229 :

$$d\gamma = \alpha(t)dt + \sigma d\epsilon \tag{6.1}$$

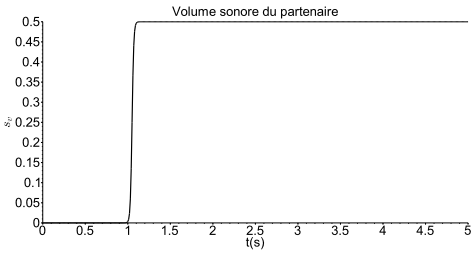
Dans l'équation 6.1, γ représente le niveau de certitude sur le comportement du partenaire (passer de locuteur à auditeur ou inversement). Selon le signe de γ l'agent est plus ou moins certain d'une alternative ou d'une autre. Lorsque $\gamma > 0$, l'agent est plus ou moins certain de la prise de tour ou de l'abandon de tour de son partenaire. Lorsque $\gamma < 0$, l'agent est plus ou moins certain que son partenaire souhaite garder son rôle. Lorsque le participant a atteint un seuil $\theta_\gamma^\pm$ en faveur d'une alternative par rapport à l'autre, l'agent est certain de la nature du comportement de son partenaire, soit changer de rôle (seuil θ_γ^+), soit garder le même rôle (seuil θ_γ^-). Dans le cas où l'agent franchit le seuil favorisant l'alternative du changement de rôle, il change lui-même automatiquement de rôle, modifiant ses équations de contrôle de l'action et la nature de son processus de perception. Dans le cas inverse, si le niveau de perception atteint la valeur θ_γ^-, l'agent réinitialise la valeur de la variable de perception à la valeur de départ 0. La réinitialisation indique que l'agent ne tient plus compte du comportement passé de son partenaire : il recommence en quelque sorte un nouveau processus de perception. Nous verrons dans le chapitre 7 que réinitialiser le niveau de certitude permet à un participant motivé à parler ou à laisser le tour de recommencer à provoquer une transition si les tentatives précédentes ont été infructueuses.

Dans notre modèle, la fonction d'accumulation $\alpha(t)$ est définie par l'équation 6.2.

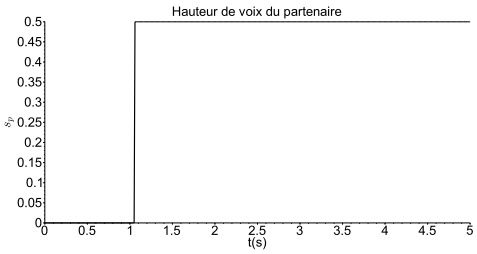
$$\alpha(t) = \sum_{j=0}^{n_s} \alpha_j(\dot{s}_j(t), s_j(t)) \tag{6.2}$$

α_j représente des fonctions d'accumulations partielles calculant le taux d'accumulation pour un signal s_j en particulier. Lorsque $\alpha_j > 0$, cela indique que le partenaire de l'agent a produit une valeur et une variation de signal s_j favorisant l'alternative d'un changement de rôle ; lorsque $\alpha_j < 0$ l'agent perçoit que son partenaire a produit une valeur et une variation d'action en défaveur d'un changement de rôle. n_s représente le nombre de signaux. α est une somme de toutes les valeurs d'accumulations α_j calculées pour chaque signal, tel que défini par l'équation 6.2.

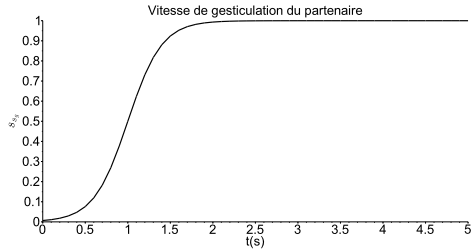
À titre d'exemple, prenons un agent locuteur capable de percevoir le volume sonore, la hauteur de voix, les regards de son partenaire et les gestes de ce dernier. En s'inspirant de l'état de l'art sur le tour de parole (voir la section 4.3 du chapitre 4) nous simulons la réalisation de quatre actions liées à la prise de parole de l'auditeur. Très simplement nous pouvons déjà considérer la présence ou non de volume sonore conjointement avec des informations de hauteur de voix comme indicateur de prise de parole. Les regards sont de même des indicateurs avérés de prise de tour (voir chapitre 4). Plus précisément nous modélisons ici la variation de la direction de



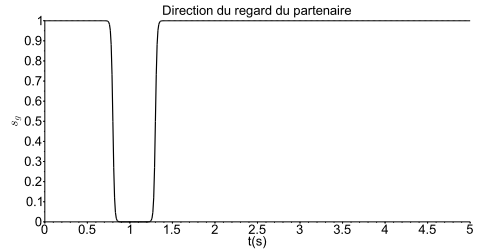
(a) Volume sonore du partenaire.



(b) Hauteur de voix du partenaire.



(c) Vitesse de geste du partenaire.



(d) Regards du partenaire.

FIGURE 6.2 – Exemple de signaux de prise de tour produits par un auditeur courant théorique.

Signal s_j	$\alpha_{s_j}(s_j, \dot{s_j})$
Volume sonore v	$\alpha_v(v, \dot{v}) = 10.0 \times (v - 0.2)$
Hauteur de voix p	$\alpha_p(p, \dot{p}) = 10.0 \times (p - 0.2)$
Vitesse de gesticulation s_g	$\alpha_{s_g}(s_g, \dot{s_g}) = 5.0 \times s_g$
Regards du participant g	$\alpha_g(g, \dot{g}) = -5.0 \times (g - 0.5)$

TABLE 6.1 – Les différentes composantes de la fonction d'accumulation α_{s_j} assignées au participant théorique

regard avec un partenaire détournant le regard juste avant de parler puis regardant de nouveau l'agent lorsqu'il a pris le tour. L'initiation d'une gesticulation est aussi vue comme un signal de prise de tour, nous proposons ici la modélisation de la vitesse de production de geste ou vitesse de gesticulation. Les variations de ces signaux sont illustrées sur la figure 6.2.

Toutes les valeurs sont théoriques et comprises entre 0 et 1. Pour le volume sonore, la hauteur de voix et la vitesse de gesticulation, une valeur de 1 correspond à une valeur maximale de ces grandeurs. Pour le volume sonore, une valeur de 0 correspond à un agent ayant arrêté de parler, tandis que pour la hauteur de voix et la vitesse de gesticulation, 0 correspond à la valeur minimale, non nulle, produite par l'agent. En ce qui concerne le regard, la valeur 1 indique un participant regardant fixement l'agent et 0 un participant ayant complètement détourné son regard de l'agent. Lorsque cette valeur est à 0.5 l'agent alterne entre des regards vers l'auditeur et des regards détournés. Nous définissons les fonctions $\alpha_{s_j}(s_j, \dot{s}_j)$ dans le tableau 6.1.

Pour le volume et la hauteur de voix, la valeur α_{s_j} a été déterminée de sorte que le taux soit négatif lorsque le volume sonore et la hauteur de voix sont à 0 et positive

lorsque ces signaux sont supérieurs à 0.2. La valeur 0.2 est une valeur arbitraire et le scénario aurait pu s'appliquer à la valeur 0.1 ou 0.3 sans changer, qualitativement, les résultats que nous allons montrer ci-dessous.

Si l'on résume le comportement du partenaire de l'agent prenant le tour, ce dernier commence à initier des gestes en augmentant lentement sa vitesse de gesticulation. Il détourne ensuite le regard au bout de 750 ms, puis prend la parole au bout d'une seconde. Nous exécutons le module de perception du comportement de l'agent de sorte de ne pas avoir pour le moment de paramètre stochastique : $\sigma = 0$ et l'équation se réduit à $d\gamma = \alpha(t)dt$. Le résultat de la perception du comportement est montré sur la figure 6.3, et l'évolution des différents taux d'accumulation partiels est détaillé sur la figure 6.4. On observe sur la figure 6.3 un taux d'accumulation négatif au départ : le partenaire n'a pas encore modulé d'actions envers la prise de tour, excepté la vitesse de gesticulation qui varie faiblement impliquant une variation lente du taux d'accumulation pour les gesticulations tel que montré sur la figure 6.4. L'agent considère que l'auditeur ne prend pas le tour. Au bout de 800 ms on observe une stabilisation du niveau de certitude de l'agent, à ce moment-là, le partenaire commence à détourner son regard, mais comme il n'a pas encore produit de son, le niveau de certitude est proche de 0. Ce n'est que lorsque le partenaire commence à parler que l'agent augmente significativement son niveau de certitude jusqu'à la valeur seuil : ici la valeur 1. On observe une courbe en dents de scie, cela est dû à l'implémentation des seuils du *DDM*. Lorsque l'agent atteint un seuil discriminant un comportement par rapport à un autre, il réinitialise son niveau de certitude γ à 0. Pour cette simulation nous avons défini que les agents ne changeaient pas de rôle, ce qui explique la succession de dents de scie après la prise de tour de l'auditeur, avec une certitude atteignant plusieurs fois le seuil positif puis étant réinitialisée. Si nous avions implémenté le changement de rôle, le fait que la variable de certitude atteigne le seuil positif aurait modifié le processus de perception de l'agent et la courbe aurait été différente.

Étudions maintenant l'effet du retrait de certains signaux sur la perception de l'agent. Actuellement l'agent atteint pour la première fois le seuil de certitude 1 au bout de 1.15 secondes.

Les figures 6.6, 6.5, 6.7, 6.8 illustrent l'effet du retrait ou l'ajout de la perception de différentes actions sur la perception de la prise de tour. On observe une variation dans le temps d'incertitude c'est-à-dire le temps requis pour atteindre pour la première fois le seuil 1 selon le signal transmis. Sans production de gestuelle l'agent atteint le seuil de certitude au bout $1,2$ secondes (voir figure 6.5), sans variation de regards l'agent a sensiblement le même temps de réaction qu'avec la variation de regard (voir figure 6.6) et sans regard ni gestuelle, l'agent atteint le seuil au bout 1,3 secondes (voir figure 6.7). Bien sûr, sans signal vocal (voir figure 6.8), aucune décision perceptive n'est prise.

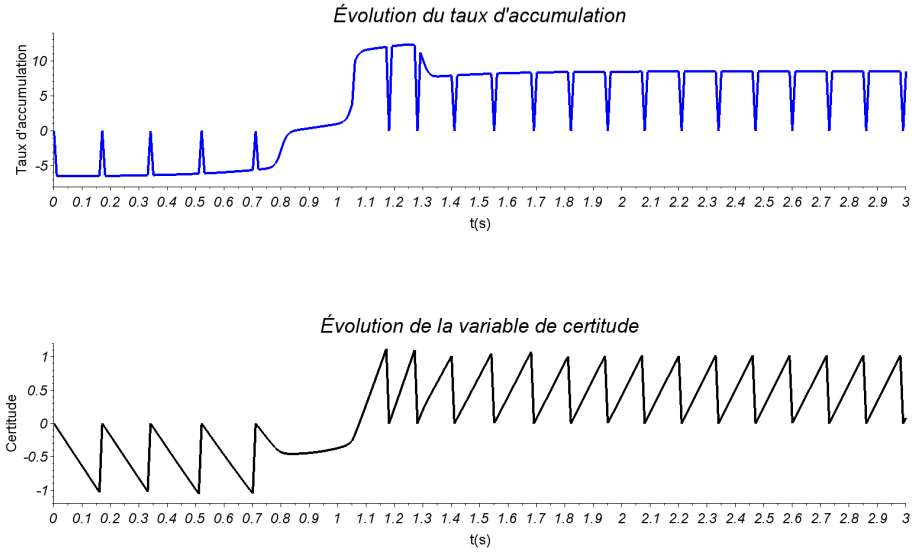


FIGURE 6.3 – Illustration de la prise de décision de l'agent lorsque tous les signaux du partenaire sont employés. Figure du haut : évolution du taux d'accumulation $\alpha(t)$. Figure du bas : variable de certitude γ.

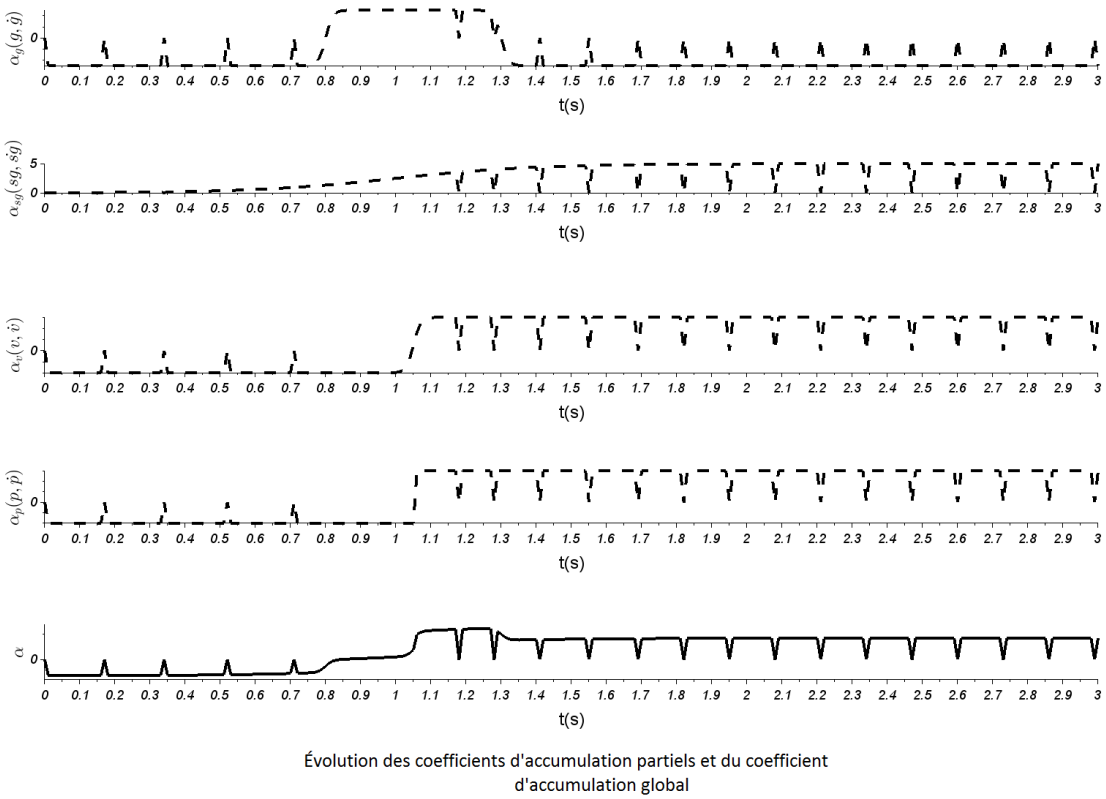


FIGURE 6.4 – Valeurs d'accumulation partielle pour les différents signaux et accumulation résultante. De haut en bas : accumulation pour le regard, la vitesse de gesticulation, le volume, la hauteur de voix (traits pointillés) et accumulation totale (trait plein).

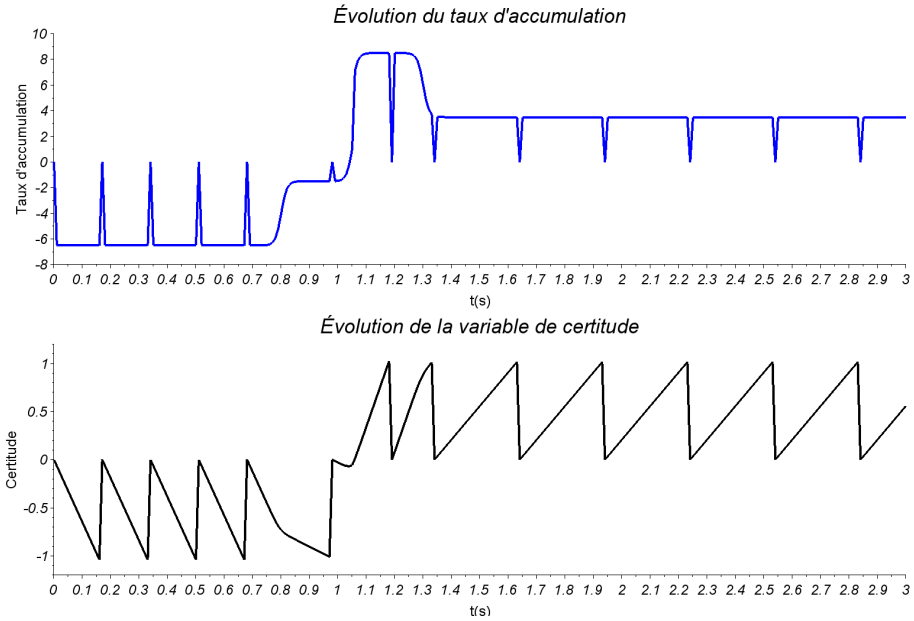


FIGURE 6.5 – Résultat de la perception du comportement en enlevant le signal de gestuelle.

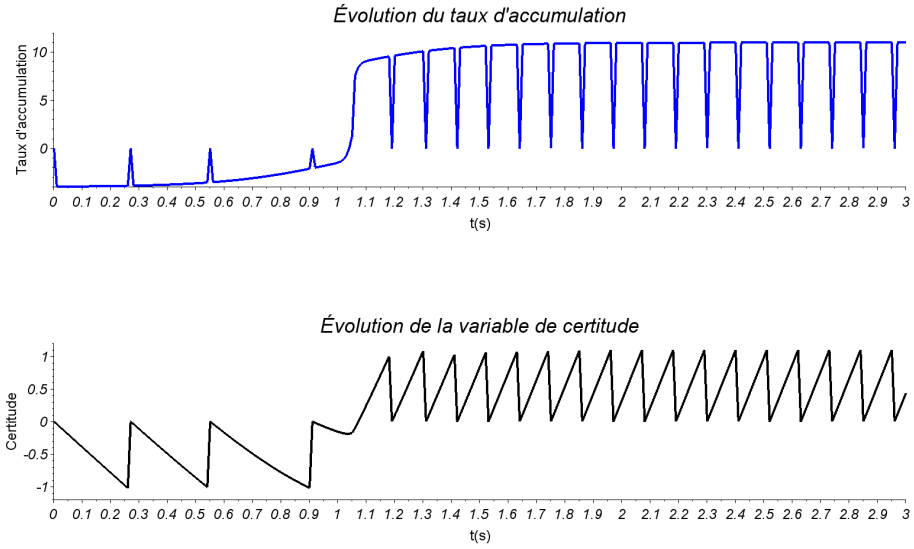


FIGURE 6.6 – Résultat de la perception du comportement en enlevant le signal de regard.

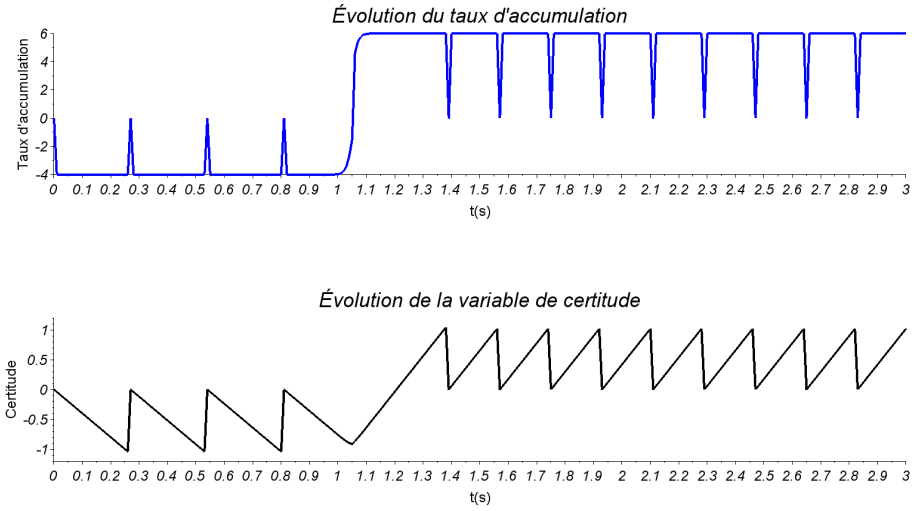


FIGURE 6.7 – Résultat de la perception du comportement en enlevant les signaux de gestuelle et de regard.

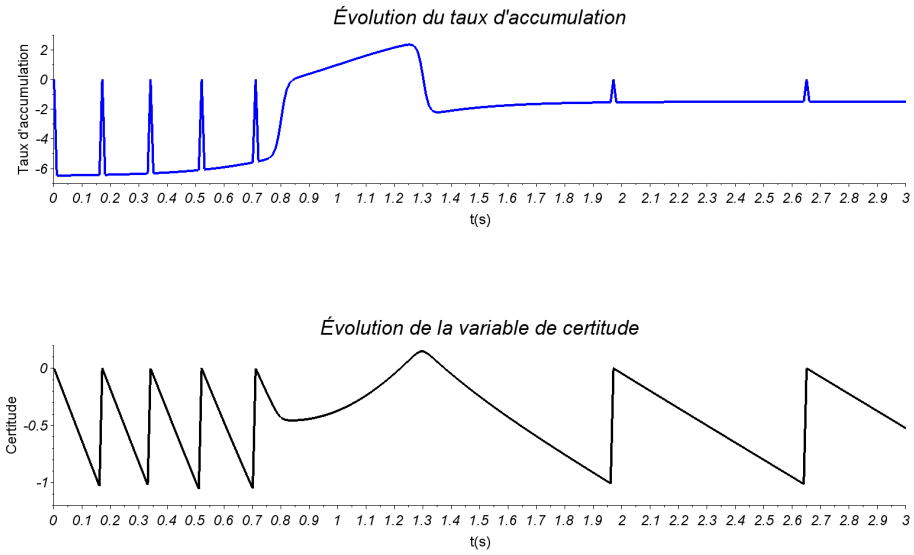
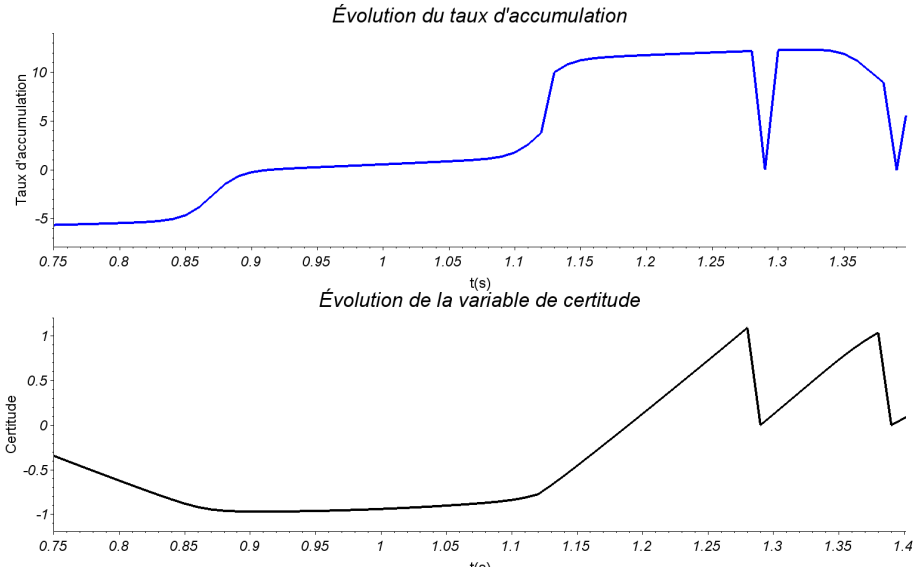


FIGURE 6.8 – Résultat de la perception du comportement en enlevant le signal de voix.

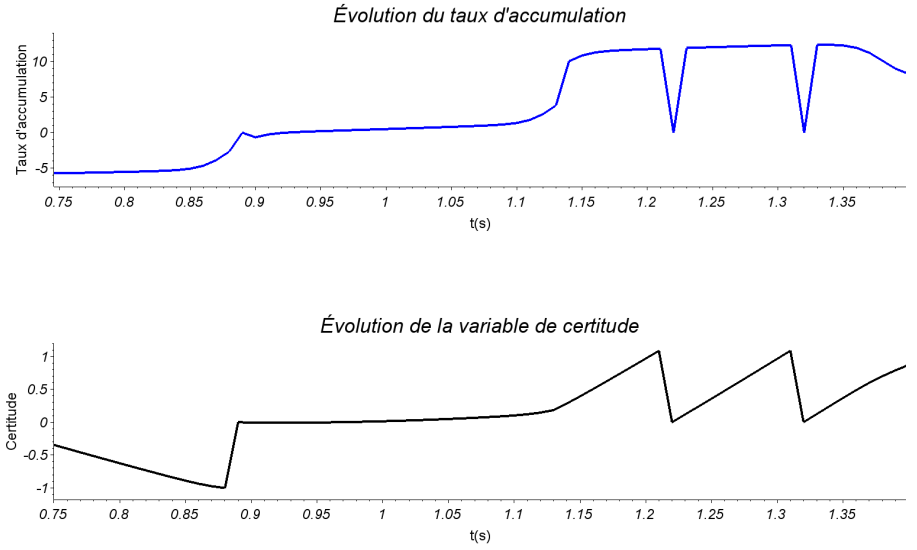
Lors de l'analyse de ces courbes, on remarque qu'une partie de la variabilité dans les prises de décision est due au niveau de certitude de l'agent avant que se produise une variation significative d'une action indiquant la prise de tour de l'agent. Si la variation de regard avait été décalée un peu plus tard, l'agent aurait été, au début de la production de la variation de regard, plutôt certain que son partenaire ne prenait pas le tour ce qui aurait résulté en une variation du temps d'incertitude. Nous proposons d'évaluer la variabilité de ce temps de prise de décision. Pour cela nous avons modifié les équations de variations des signaux illustrés sur la figure 6.2 de sorte d'introduire un décalage dans la production de signal. Par ce décalage, les variations de signaux liés à la prise de tour coïncident avec différents moments dans le processus de perception de l'agent. La figure 6.9 montre ainsi deux simulations effectuées pour deux valeurs temporelles de décalage, une à 0.07 seconde, impliquant que l'agent commence à produire sa variation de regard au bout de 860 ms au lieu de 790 ms et une valeur de décalage de 0.08 secondes. Lorsque la valeur de décalage est à 0.07, l'agent est proche d'atteindre le seuil négatif de perception avec un niveau de certitude proche de -1 (figure 6.9a) lorsque le partenaire commence à produire le détournement de son regard, l'agent vient de franchir le seuil négatif de perception, résultant en une valeur initiale de 0 (figure 6.9b). Ces deux cas de figure laissent une variabilité dans la prise de décision de 70 ms au maximum. Cette variabilité peut être minimisée en augmentant les taux d'accumulations. La dérivée de la variable de certitude sera alors plus grande, minimisant le temps qu'il faut à la valeur de certitude pour atteindre la valeur seuil et diminuant l'effet de la valeur de certitude de départ sur le temps de perception.

Nous simulons maintenant le processus de perception du comportement avec un écart-type non nul tel que spécifié par le modèle de dérive-diffusion (*cf.* Équation B.1). Beaucoup de facteurs environnementaux peuvent influer sur les processus cognitifs du participant, sans compter que la perception et l'action d'un participant humain dépendent de l'historique de son interaction avec l'environnement, résultant en une variabilité inter-participants dans le processus de décision. De Ruiter *et al.* (2006) montrent que moins le nombre de signaux auxquels ont accès les participants est grand, plus la variance dans l'estimation du temps de fin de tour est grande. Nous souhaitons valider ce principe en vérifiant que plus le nombre de signaux ajoutés est grand moins l'écart-type dans le temps d'incertitude (ici le temps où la certitude atteint pour la première fois la valeur 1) est grand. Nous avons fixé le paramètre d'écart-type σ de l'équation 6.1 de perception du comportement à 20.0 soit une valeur d'écart-type grande. Nous montrons sur la figure 6.10 un exemple d'une perception dégradée avec cet écart-type.

Nous avons effectué 1000 simulations pour chaque sous-ensemble de signaux, la figure 6.11 montre les résultats obtenus. On remarque que malgré un bruit stochastique élevé dans le processus de décision, lorsque l'agent interprète tous les signaux,



(a) Prise de décision où le niveau de certitude est au plus bas lorsque son partenaire détourne son regard



(b) Prise de décision où le niveau de certitude est au plus haut lorsque son partenaire détourne son regard

FIGURE 6.9 – Figures illustrant la variabilité du temps de perception de la prise de tour en fonction des conditions initiales

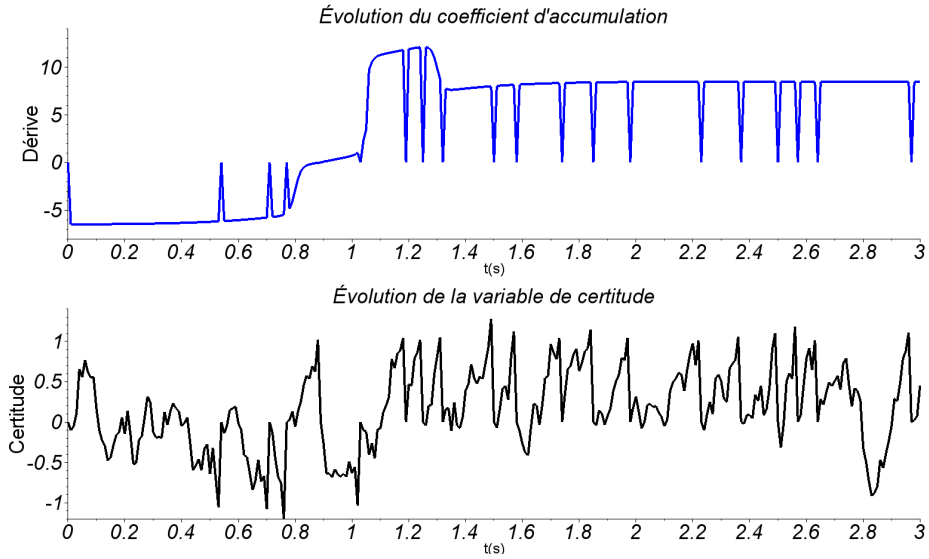


FIGURE 6.10 – Illustration d'un processus de prise de décision lorsque le terme stochastique est non nul et le paramètre σ est fixé à 20.0.

il parvient à estimer la prise de tour de l'agent avec une variance faible (intervalle inter-quartile $iqr = 0.19$). Lorsque l'on enlève la gestuelle on modifie peu la variance de l'estimation de la prise de tour ($iqr = 0.17$). Ces deux sous-ensembles de signaux se distinguent nettement du reste des signaux. En effet retirer les regards double la variance ($iqr = 0.47$), le même effet étant observé lorsque l'on garde uniquement les signaux non-vocaux ($iqr = 0.47$) et lorsque l'on garde uniquement les signaux para-verbaux ($iqr = 0.54$). Les valeurs médianes des temps de prise de décision sont les mêmes peu importe les signaux employés par l'agent. La similarité dans ces valeurs médianes proviennent du fait que les valeurs d'accumulation pour chaque simulation sont les plus élevées aux alentours de 1.2 s conduisant en moyenne à une prise de décision au même endroit peu importe l'ensemble de signaux employés.

Ces simulations illustrent bien un principe général d'additivité des signaux, reproduisant les résultats de certaines études (Gravano et Hirschberg, 2011; Hjalmarsson, 2011). L'interprétation des gesticulations a peu d'effets sur la prise de décision de l'agent. Au contraire, l'enlever semble améliorer la prise de décision. Cela est dû au fait que les signaux de gesticulation sont détectés comme des signaux de prise de tour bien avant les autres signaux. La valeur de certitude est donc légèrement supérieure au début de la simulation lorsque les signaux de gesticulation sont présents, augmentant la probabilité de franchir le seuil positif de décision plus tôt. Il est à noter que ce résultat, lié à l'effet peu important des gesticulations sur la prise de décision n'est vrai que dans notre scénario théorique mais ne correspond pas forcément à ce qu'il se passe réellement dans des conversations humaines.

Bien sûr si l'on augmente encore le paramètre stochastique nous finissons par avoir une prise de décision entièrement bruitée. Ainsi si l'on augmente la valeur du paramètre σ à 40.0 on observe que la répartition reproduit les caractéristiques d'une

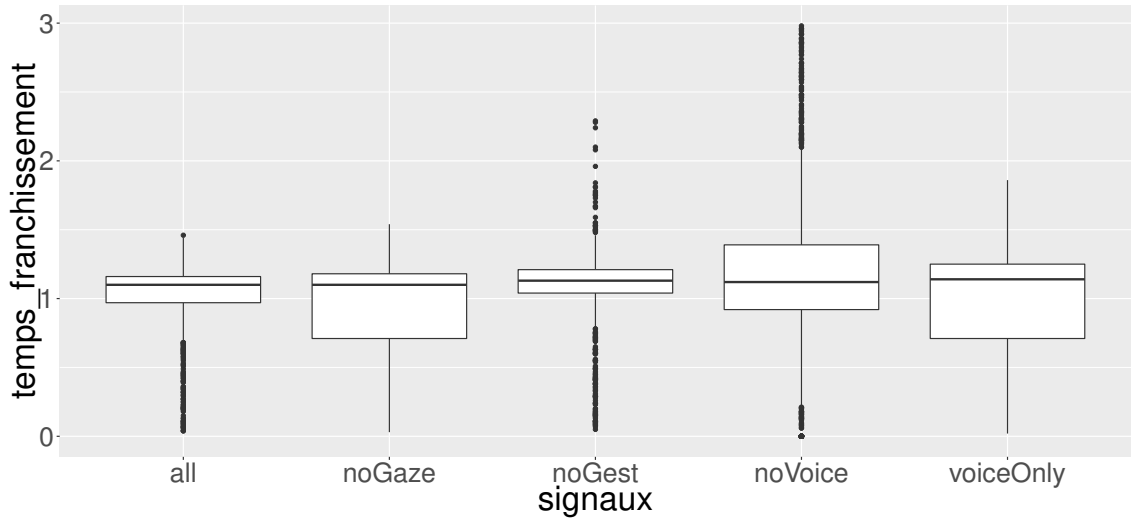


FIGURE 6.11 – Diagrammes en boite des résultats de la simulation pour différents signaux.

loi de probabilité exponentielle impliquant que les variations de signaux ont peu d'effets sur la prise de décision de l'agent qui est alors entièrement aléatoire.

Nous avons détaillé ici le fonctionnement du module de perception du comportement de l'utilisateur. Nous allons maintenant nous intéresser à la formulation générale des équations de modulation des actions.

6.2.3 Modulation des actions

Suivant les principes de la dynamique comportementale, chaque action de l'agent est contrôlée par un système dynamique. L'ensemble des systèmes dynamiques contrôlant la production des actions de l'agent représente les lois de contrôle de l'agent. En conformité avec les hypothèses 1 et 7 la motivation à parler m et le niveau de certitude du partenaire γ sont des variables de contrôle du modèle, illustrant une influence directe de la perception du comportement de l'utilisateur, donc de l'information provenant de l'environnement, sur le comportement de l'agent. En conformité avec l'hypothèse 1, la motivation m est liée à un but de l'agent : garder ou abandonner le tour s'il est locuteur ou prendre ou ne pas prendre le tour s'il est auditeur.

Application des systèmes dynamiques au contrôle des actions de l'agent

Dans notre modèle, tous les systèmes dynamiques de contrôle des actions de l'agent ont la même forme générale, donnée par l'équation 6.3 :

$$\ddot{a_j} = -b \times \dot{a_j} - k_g \times (a_j - f(m, \gamma)) \tag{6.3}$$

Les variables d'action dans le modèle théorique sont formulées de manière abstraite : chaque action j produite par l'agent varie dans l'intervalle $[0.0, 1.0]$, 0.0 représentant

la valeur minimale de l'action et 1.0 la valeur maximale de l'action. La manière dont ces « actions » théoriques sont converties en grandeurs réelles dépend de la nature de chaque signal et de l'actionneur chargé d'exécuter l'action correspondant au signal produit. La motivation m est représentée par une valeur variant entre -1 représentant une forte motivation de ne pas changer de rôle de la part de l'agent et 1 représentant une forte motivation de changer de rôle.

$-b \times \dot{a_j}$ est un terme d'amortissement correspondant à une inertie intrinsèque à l'agent dans la production de signal. Le terme $-k_g \times (a_j - f(m, \gamma))$ détermine vers quelle valeur le signal varie en fonction de la valeur d'accumulation actuelle de l'agent et sa propre motivation. k_g représente le paramètre de raideur de l'équation, plus ce paramètre est grand plus l'attraction ou la répulsion du point fixe dans l'espace d'état sera grand. Illustrons l'influence de la raideur et de l'amortissement sur un exemple simple.

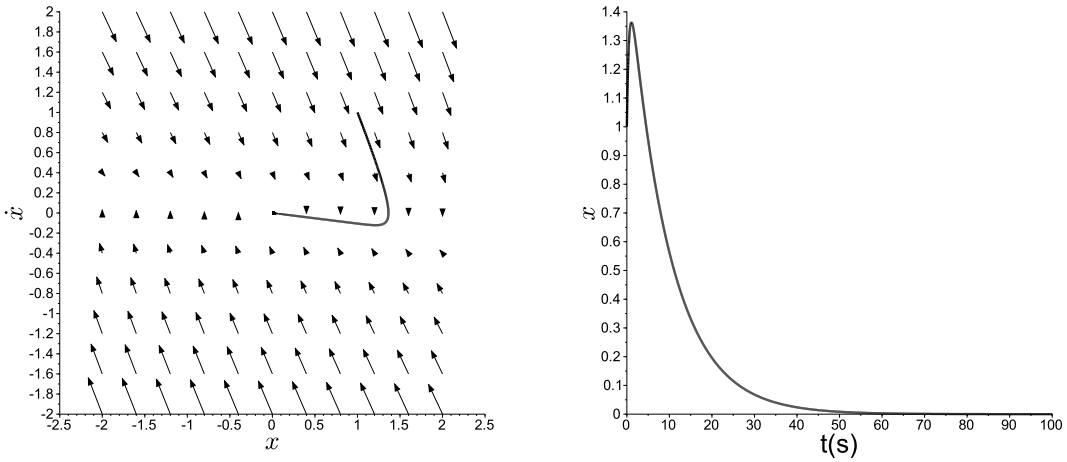
Pour la modélisation du tour de parole nous avons choisi de nous placer systématiquement en régime apériodique, nous définissons donc $k_g < \frac{(b^2)}{4}$. Nous pouvons modifier la valeur de b ou de k_g pour obtenir un agent ayant une inertie plus ou moins forte et rendre l'attracteur plus ou moins stable aux perturbations. Les figures 6.12 et 6.13 montrent l'effet d'un paramètre d'amortissement b plus petit et d'un paramètre de raideur k_g plus grand. On observe dans les deux cas une convergence plus rapide vers l'attracteur du système.

$f(m, \gamma)$ est l'attracteur du système représentant la valeur vers laquelle converge le signal. Lorsque cette fonction varie, l'attracteur du système varie selon l'axe des x dans l'espace d'états du système. L'attracteur varie ici de sorte que la valeur de l'action de l'agent converge vers la valeur courante de $f(m, \gamma)$. Les figures 6.14a et 6.14b montrent le résultat d'une simulation pour deux attracteurs différents, l'un à 0.5 et l'autre à 1.0.

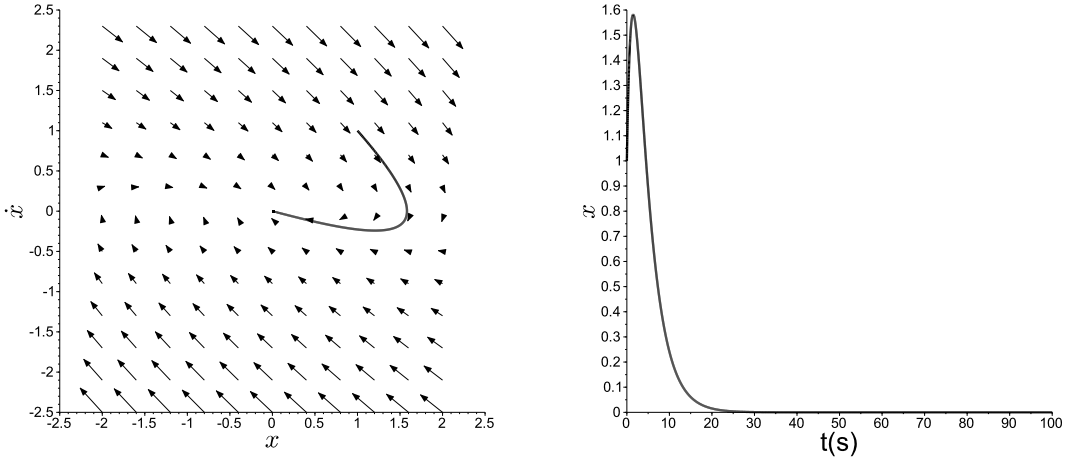
Dans notre modèle, $f(m, \gamma)$ est dynamique dépendant à l'instant t de la motivation m de l'agent et du niveau de certitude γ. Nous définissons dans la suite la fonction sigmoïde $S(x)$ selon l'équation 6.4. La figure 6.15 est une représentation graphique de la fonction sur l'intervalle $[-10, 10]$.

$$S(x) = \frac{1}{1 + \exp(-5 \times x)} \tag{6.4}$$

Illustrons le cas d'un agent modulant son volume sonore, sa hauteur de voix et son débit de parole. Lorsque $m < 0$ et $\gamma > 0$ le volume sonore et la hauteur de voix augmentent représentant une situation de conflit, et le débit de parole diminue en accord avec les observations effectuées par Schegloff (2000). Lorsque $m > 0$ et $\gamma > 0$ le volume sonore, la hauteur de voix et le débit de parole diminuent jusqu'à la valeur 0. Nous définissons les fonctions $f_p(m, \gamma)$, $f_v(m, \gamma)$ et $f_r(m, \gamma)$ selon l'équation 6.5.

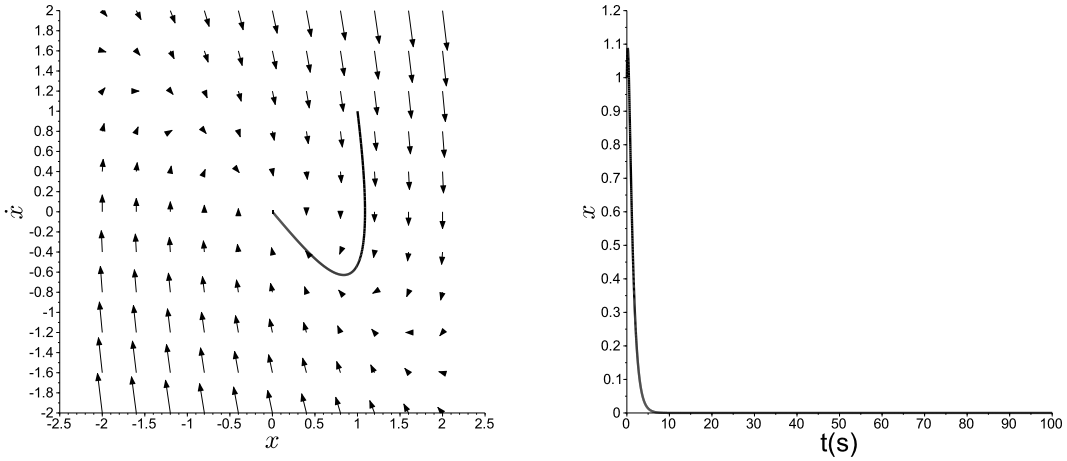


(a) Simulation de l'équation A.1 pour $b = 2$.

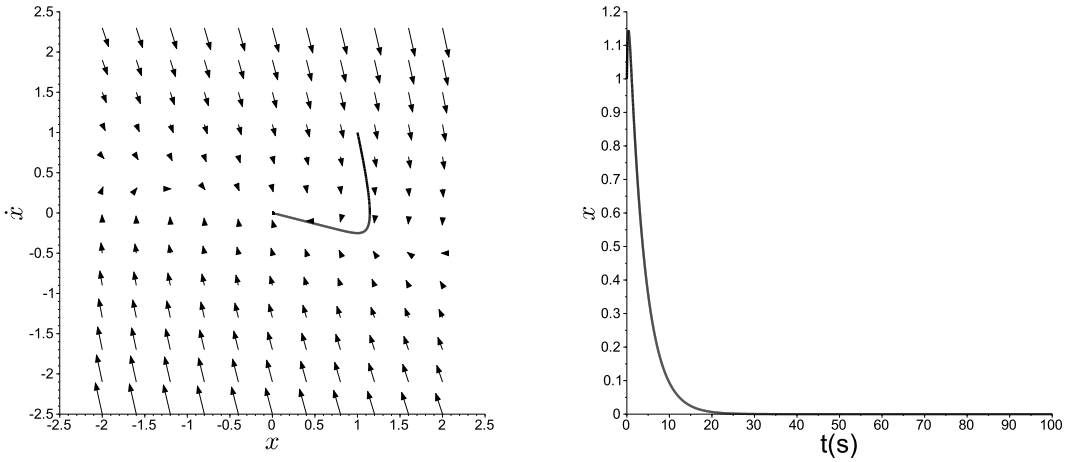


(b) Simulation de l'équation A.1 pour $b = 1$.

FIGURE 6.12 – Effet de la variation de b sur le comportement du système. Pour chaque figure, à gauche est représenté l'espace d'états de l'équation sous la forme d'un champ de vecteur, et la trajectoire représentant l'évolution de l'état du système (trait plein). À droite est représentée l'évolution temporelle du système.

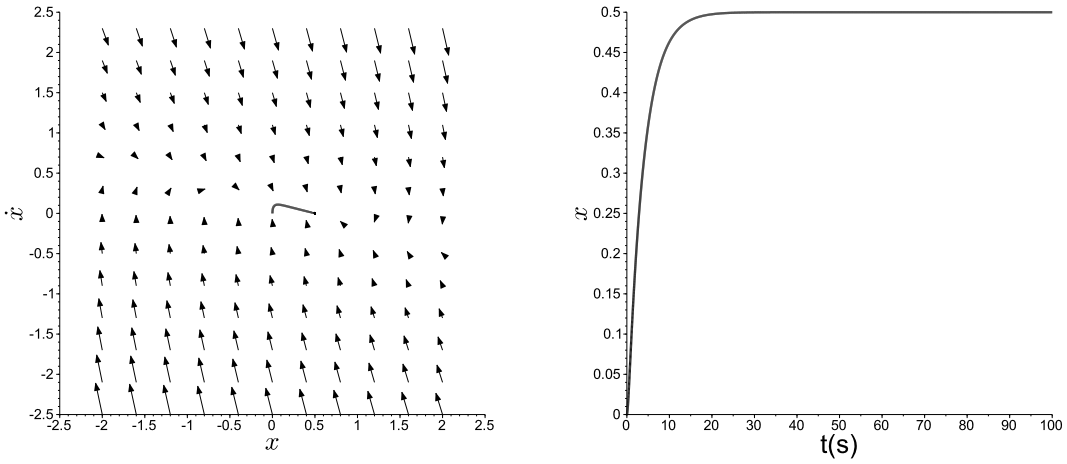


(a) Simulation de l'équation A.1 pour $k_g = 3$.

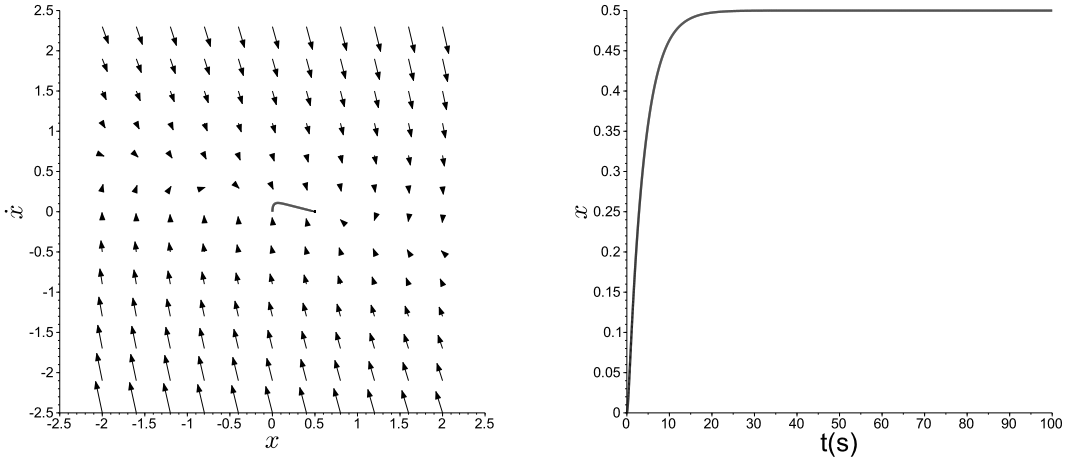


(b) Simulation de l'équation A.1 pour $k_g = 1$.

FIGURE 6.13 – Effet de la variation de k_g sur le comportement du système.



(a) Trajectoire résultant de la simulation de l'équation différentielle définie par l'équation 6.3 avec $f(m, \gamma) = 0.5$.



(b) Trajectoire résultant de la simulation de l'équation différentielle définie par l'équation 6.3 avec $f(m, \gamma) = 1.0$.

FIGURE 6.14 – Exemples de simulations de l'équation 6.3 pour deux valeurs de $f(m, \gamma)$.

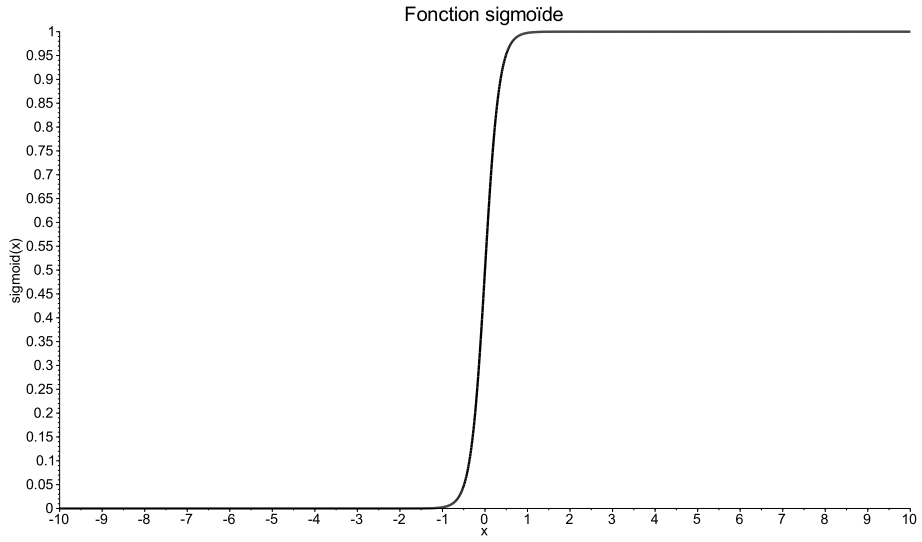


FIGURE 6.15 – Représentation de la fonction $S(x)$ sur l'intervalle $[-10, 10]$.

$$\begin{cases} f_p(m,\gamma) = f_v(m,\gamma) = 0.5 + 0.5 \times S(\gamma) \times (S(-m) - S(m)) \\ f_r(m,\gamma) = 0.5 - 0.5 \times S(\gamma) \end{cases} \tag{6.5}$$

$f_p(m,\gamma)$ et $f_v(m,\gamma)$ sont identiques. Il s'agit d'une approximation provenant du fait que la variation de hauteur de voix et de volume sonore suivent la même dynamique selon les signes de m et γ. Les fonctions ont deux termes, 0.5 représente une valeur moyenne de l'action de l'agent, la valeur par défaut lorsque l'agent parle. Lorsque $\gamma < 0$ le second terme est négligeable, l'agent continue donc de parler avec un niveau sonore égal à 0.5. Lorsque $\gamma > 0$ l'évolution de l'action de l'agent dépend de la motivation m. Une valeur $m < 0$ augmentera l'attracteur de l'action jusqu'à la valeur 1, le terme $S(m)$ étant négligeable par rapport à $S(-m)$, une valeur $m > 0$ diminuera l'attracteur de l'action à 0, le terme $S(-m)$ étant négligeable par rapport à $S(m)$. Le débit de parole diminuant seulement lorsque $\gamma > 0$, la fonction $f_r(I,\gamma)$ est définie de sorte que le débit de parole diminue lorsque $\gamma > 0$ et reste constant à 0.5 lorsque $\gamma < 0$.

Nous définissons l'amortissement de l'équation de contrôle du volume sonore $b_v = 10.0$ et sa raideur $k_{g_v} = 20.0$, l'amortissement de la hauteur de voix $b_p = 5.0$ et sa raideur $k_{g_p} = 5.0$ et l'amortissement de la variation du débit de parole $b_r = 2.0$ et sa raideur $k_g = 1.0$. Ces valeurs sont purement théoriques et ne correspondent à aucune donnée réelle provenant d'interactions humaines.

Nous simulons ici une croissance de γ telle que $\gamma(t) = 1 - 2 * \exp(-t)$ avec $t \in [0.0, 10.0]$. Le temps t de la simulation est un temps théorique n'ayant pas d'unité attribuée. La figure 6.16 montre les résultats de la simulation.

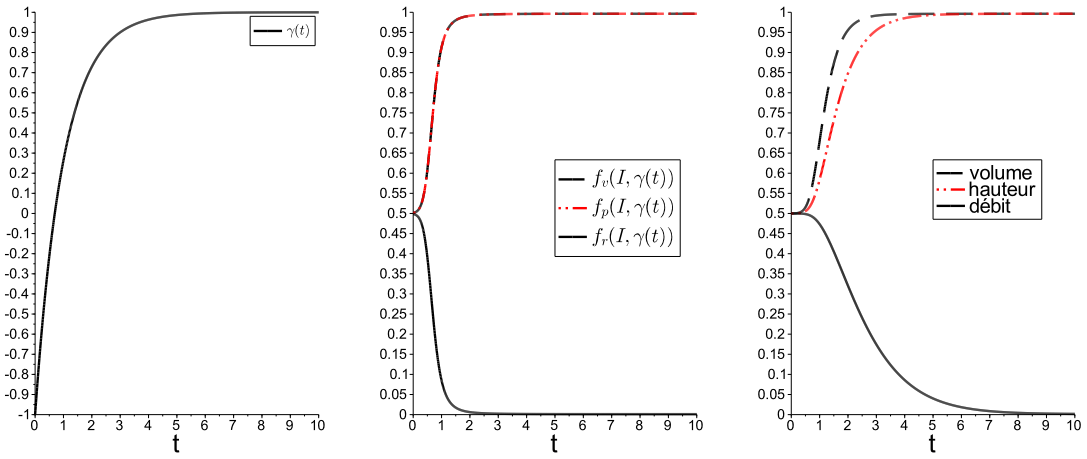


FIGURE 6.16 – Illustration d'une simulation avec une variation du niveau de certitude γ de -1 à 1 suivant l'équation $\gamma(t) = 1 - 2 * \exp(-t)$ comme illustré sur la figure de gauche. La variation de $f_v(m,\gamma)$, $f_p(m,\gamma)$, $f_r(m,\gamma)$ est illustrée sur la figure du milieu et la variation des actions sur la figure de droite.

Dans cette simulation $\gamma = -1$ à $t = 0$, l'agent commence donc avec des valeurs d'actions à 0.5. À mesure que le degré de certitude augmente, les attracteurs du volume et de la hauteur commencent à augmenter tandis que l'attracteur du débit de parole diminue, résultant en une croissance du volume et de la hauteur de voix et une décroissance du débit de parole. L'attracteur se stabilise enfin à 1 pour $f_v(m,\gamma)$ et $f_p(m,\gamma)$ résultant quelque temps plus tard en une convergence du volume sonore et de la hauteur de voix vers la valeur 1. $f_r(m,\gamma)$ se stabilise en 0 résultant en une convergence du débit de parole vers 0.

Si nous reproduisons la même simulation pour une motivation de laisser la parole de 1.0, nous obtenons les résultats présentés sur la figure 6.17, simulant un agent baissant son volume sonore, sa hauteur de voix et son débit de parole pour laisser son partenaire prendre le tour.

Dans les deux cas, les variables de contrôle γ et m contraignent la variation des actions de l'agent de sorte qu'une synergie se crée entre ces différentes actions, en accord avec l'hypothèse 4.

6.3 Conclusion

Nous avons présenté dans ce chapitre nos hypothèses concernant les propriétés d'un modèle de tour de parole puis nous avons présenté le fonctionnement du modèle général et nous avons détaillé ses composantes. Nous avons illustré d'une part la capacité d'un agent à percevoir de manière fiable le comportement de son partenaire malgré une décision très bruitée. D'autre part nous avons montré sur un exemple comment l'agent pouvait synchroniser ses actions en vue de prendre ou abandonner

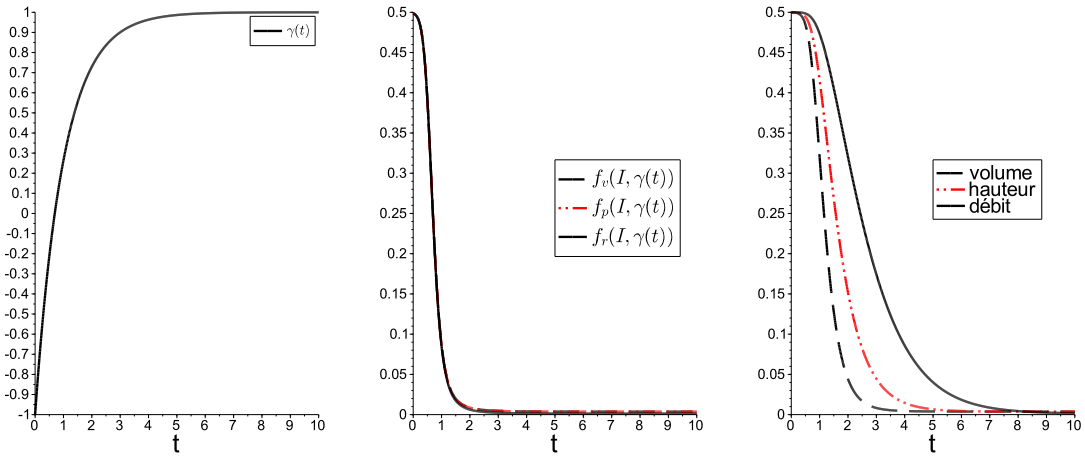


FIGURE 6.17 – Illustration d'une simulation avec une variation du niveau de certitude γ de -1 à 1 selon l'équation $\gamma(t) = 1 - 2 * \exp(-t)$, comme illustré sur la figure de gauche. La variation de $f_v(m,\gamma)$, $f_p(m,\gamma)$, $f_r(m,\gamma)$ est illustrée sur la figure du milieu et la variation des actions sur la figure de droite.

le tour par l'influence directe de sa motivation et de son niveau de certitude sur la nature du comportement de son partenaire. Nos hypothèses ont de plus bien été implémentées par ce modèle. L'hypothèse 1 est modélisée par la variation de la motivation à parler de l'agent pouvant provenir du traitement du contenu du dialogue, de l'attitude de l'agent envers son partenaire ou d'autres variables contextuelles. La présence de processus de perception et d'action, peu importe le rôle du participant, implique la capacité des agents à agir lorsqu'ils sont auditeur et à percevoir lorsqu'ils sont locuteur. L'auditeur peut ainsi être à l'initiative d'un changement de tour, en accord avec l'hypothèse 2. L'hypothèse 3 correspond dans notre modèle au principe même du module de prise de décision. L'agent interprète et fusionne l'ensemble des informations provenant des actions qu'il peut interpréter. C'est la valeur de certitude de l'agent qui influe sur la production de toutes les actions, et non pas un sous-ensemble d'actions. De la même manière, l'ensemble des actions de l'agent est contrôlé par les mêmes variables : la certitude γ et la motivation m, plutôt que par des variables distinctes selon l'action produite par l'agent (hypothèse 4). Le principe d'accumulation des actions (hypothèse 5) est illustré dans notre modèle par l'implémentation du *DDM*, la valeur de certitude γ résulte alors de ce processus d'accumulation et influence directement la production des actions de l'agent en accord avec l'hypothèse 7. L'hypothèse 6 est illustrée plus en détail dans le chapitre 7.

Chapitre 7

Analyse du comportement du modèle

Dans le chapitre 6 nous avons présenté séparément les deux composantes principales de notre modèle, d'une part, la perception du comportement du partenaire et d'autre part, le contrôle de l'action de l'agent. Nous proposons dans ce chapitre une analyse du fonctionnement du modèle lorsque deux agents interagissent l'un avec l'autre. Nous illustrons ici le caractère émergent du comportement des deux agents, et leur couplage dans la modulation de leurs actions. C'est à cette condition qu'une coordination peut avoir lieu entre les deux agents. Suivant d'autres approches situées de l'interaction humaine (De Loor *et al.*, 2009; Ikegami et Iizuka, 2007), nous montrons de plus que les conditions de couplage entre les deux agents conduisent à une adaptabilité de l'agent au partenaire qu'il a en face de lui et à une robustesse à différentes conditions environnementales.

7.1 Présentation de l'implémentation

Nous présentons ici les équations utilisées dans la suite de ce chapitre pour illustrer les propriétés de notre modèle en interaction avec un partenaire artificiel. Les agents modulent trois variables, le volume sonore, la hauteur de voix et la direction de leur regard. Les valeurs prises par ces variables et leur signification provient directement du chapitre 6. Un volume sonore à 0 indique un agent qui a arrêté de parler, un volume sonore à 1 correspond à un niveau sonore élevé. La valeur moyenne de la hauteur de voix est, elle, à 0.5, la valeur 0 correspond à la hauteur de voix minimale et non nulle de l'agent, la valeur 1 correspond à la hauteur de voix maximale. Pour le regard, une valeur de 1 indique que l'agent regarde fixement son partenaire, 0.5 indique que l'agent varie sa direction de regard de sorte que 50% du temps il ne regarde pas son partenaire et 50% du temps il regarde son partenaire et 0 indique que l'agent ne regarde jamais son partenaire. Nous définissons pour chaque action les fonctions d'accumulation présentées dans le tableau 7.1.

Grandeur	Locuteur	auditeur
Volume	$\alpha_{vloc}(v, \dot{v}) = 1.5 \times (v - 0.1)$	$\alpha_{vlis}(v, \dot{v}) = -2.0 \times (v - 0.4)$
Hauteur de voix	$\alpha_{ploc}(p, \dot{p}) = 1.5 \times (p - 0.1)$	$\alpha_{plis}(p, \dot{p}) = -2.0 \times (p - 0.4)$
Regards du participant	$\alpha_{gloc}(g, \dot{g}) = -1.5 \times (g - 0.5)$	$\alpha_{glis}(g, \dot{g}) = 1.0 \times (g - 0.5)$

TABLE 7.1 – Les différentes fonctions d'accumulation partielle α_{s_j} assignées au participant selon son rôle.

Nous avons déterminé les fonctions d'accumulation de sorte d'interpréter des variations de signaux similaires à ce que l'on peut observer dans les interactions humaines (voir section 4.3 du chapitre 4, page 66). Par exemple, une variation de volume sonore négative, corrélée à une fin de tour dans un grand nombre d'études sur les interactions humaines, générera une accumulation positive pour un agent auditeur piloté par notre modèle. Nous nous intéressons ici uniquement au signes des variations de signaux et ne cherchons pas à reproduire de manière quantitative les variations de signaux. Pour le locuteur, les équations présentées dans le tableau 7.1 ont la même forme que celles présentées dans le tableau 6.1 du chapitre 6. Ainsi, un volume sonore et une hauteur de voix de l'auditeur supérieurs à 0.1 conduisent à une valeur d'accumulation positive. De la même manière, une valeur de direction du regard inférieure à 0.5 signale un auditeur détournant son regard pour prendre le tour, la valeur d'accumulation résultante est donc positive. Au contraire un auditeur regardant fixement le locuteur résulte en une accumulation négative.

Pour l'auditeur, la direction du regard correspond à l'opposé de la fonction d'accumulation de la direction du regard du locuteur : lorsque le locuteur regarde fixement l'auditeur, ce dernier l'interprète comme un signal d'abandon de tour du locuteur, la valeur d'accumulation est donc logiquement positive. Pour le volume sonore et la hauteur de voix, deux cas de figure sont à distinguer. D'une part, l'abandon de tour étant marqué par une baisse du volume sonore et de la hauteur de voix, il est logique de considérer qu'en dessous d'un certain seuil le volume sonore du partenaire marque l'abandon de tour. Nous fixons ce seuil à 0.4, de telle sorte qu'en dessous de cette valeur, la valeur d'accumulation devient positive. À l'inverse, lorsque le locuteur courant signifie qu'il ne souhaite pas laisser le tour à l'auditeur courant, il augmente son niveau sonore. Aussi la valeur d'accumulation est négative lorsque le volume sonore est supérieur à 0.4. L'accumulation de la hauteur de voix est définie par la même fonction que le volume sonore.

Pour le contrôle des signaux de l'agent, nous proposons de construire ces équations à partir de la fonction sigmoïde S présentée dans le chapitre 6 (équation 6.4). Cette sigmoïde est utilisée pour définir, selon la relation entre la certitude γ et la motivation m, l'espace des attracteurs pour chaque équation. Chaque fonction comporte plusieurs termes de la forme $f(\gamma, m) \times S(g(m, \gamma)) \times S(h(m)) \times S(v(\gamma))$. La multiplication par une fonction sigmoïde permet de définir des valeurs non nulles dans la zone de l'espace où $g(m, \gamma) > 0$, $h(m) > 0$ et $v(\gamma) > 0$. Si l'une des valeurs

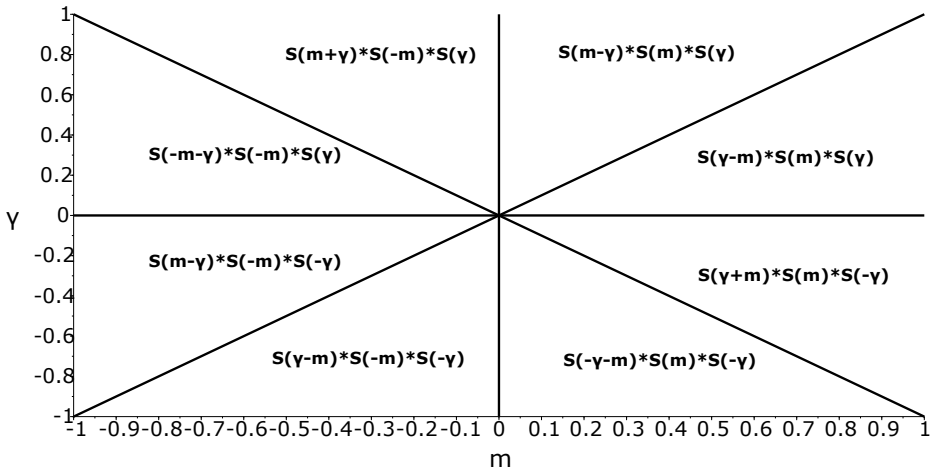


FIGURE 7.1 – Illustration de la segmentation de l'espace d'états des attracteurs selon les signes de m et γ. En abscisse est représentée m, en ordonnée est représentée γ.

résultantes de ces fonction est inférieure à 0 la fonction sigmoïde correspondante devient nulle et le terme tout entier s'annule. $S(g(m,\gamma)) \times S(h(m)) \times S(v(\gamma))$ permet ainsi de ne définir des valeurs que pour la zone où la sigmoïde est non nulle. Nous posons $g(m,\gamma)$ de la forme $\pm m \pm \gamma$, $h(m)$ de la forme $\pm m$ et $v(\gamma)$ de la forme $\pm\gamma$. Selon les différents signes que peuvent prendre g, h et v, l'espace des attracteurs peut ainsi être segmenté en huit zones, présentées sur la figure 7.1.

Pour chaque rôle, nous décrivons maintenant pas à pas la conception des différentes équations de contrôle des actions.

7.1.1 Actions du locuteur

L'agent contrôlant son niveau sonore et sa hauteur de voix de la même manière, l'espace des attracteurs pour les deux équations de contrôle correspondantes sera défini par la même équation.

Pour le locuteur nous proposons une valeur par défaut du volume sonore et de hauteur de voix à 0.5. Posons donc pour l'instant $v_{loc} = p_{loc} = 0.5$.

Lorsque $m < 0$ et $\gamma < 0$ la valeur des signaux est égale à 0.5, par contre $m < 0$ et $\gamma > 0$ indique un conflit : l'agent a pour but de garder la parole alors qu'il a un niveau de certitude indiquant que l'auditeur change de rôle. L'agent va donc augmenter son niveau sonore et sa hauteur de voix tel qu'observé dans les interactions humaines par Kurtić *et al.* (2013). L'agent module alors l'intensité avec laquelle il augmente ses signaux selon m et γ. Complétons alors les fonctions v_{loc} et p_{loc} tel que $v_{loc} = p_{loc} = 0.5 - 0.5 \times \gamma \times m \times S(-m) \times S(\gamma)$. Une motivation négative et faible et une certitude positive et faible donnera une augmentation faible des signaux. Plus la certitude ou la motivation augmente, plus la valeur de l'attracteur augmente et plus l'agent augmente la valeur de ses signaux.

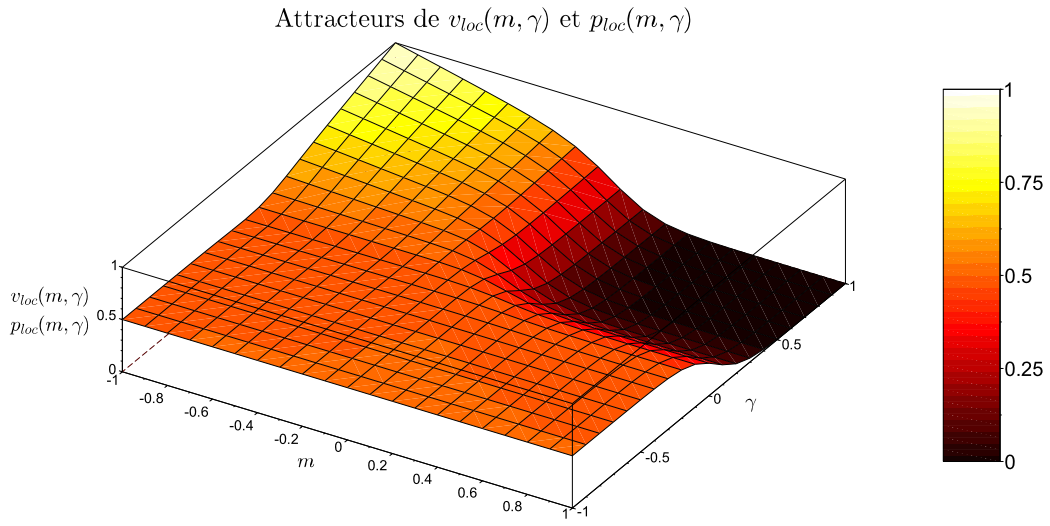


FIGURE 7.2 – Attracteurs de $v_{loc}(m,\gamma)$ et $p_{loc}(m,\gamma)$.

Lorsque $m > 0$ et $\gamma > 0$ l'agent est plus ou moins certain que l'auditeur est en train de prendre le tour. Puisqu'il souhaite laisser le tour, il diminue son volume sonore et sa hauteur de voix en accord avec les observations faites dans les interactions humaines par Gravano et Hirschberg (2011). La diminution de la valeur des signaux est provoquée par un attracteur à 0 dans l'équation de contrôle. Nous ajoutons donc un troisième terme à v_{loc} et p_{loc} tel que la fonction soit définie par l'équation 7.1. L'espace des attracteurs pour l'ensemble des valeurs de m et γ par l'équation 7.1 est illustré sur la figure 7.2.

$$\begin{aligned} v_{loc} \quad = p_{loc} \quad &= 0.5 \\ &-0.5 \times \gamma \times m \times S(-m) \times S(\gamma) \\ &-0.5 \times S(\gamma) \times S(m) \end{aligned} \tag{7.1}$$

Pour la direction du regard du locuteur courant, nous définissons comme valeur par défaut 0.5 : l'agent varie la direction de son regard en alternant les regards vers l'auditeur avec des regards détournés. Posons donc $g_{loc} = 0.5$. Le locuteur aura tendance à regarder plus fixement son auditeur pour lui laisser le tour selon les observations effectuées par Novick *et al.* (1996) pour les interactions humaines. Nous proposons alors que, lorsque $m > 0$ et $\gamma > 0$, la valeur de la direction du regard augmente. Cette augmentation est pondérée par m et γ tels que lorsque m et γ sont proches de 0 l'agent varie peu sa direction de regard, et plus m ou γ augmente, plus le regard de l'agent sera fixé sur l'auditeur courant. Nous modifions alors la fonction g_{loc} de sorte que $g_{loc} = 0.5 + 0.5 \times \gamma \times m \times S(m) \times S(\gamma)$. À l'inverse lorsque $m > 0$ et $\gamma < 0$, l'agent essaiera de provoquer le changement de tour en regardant son auditeur. Nous définissons ainsi que lorsque $\gamma < 0$ et $m > 0$, plus l'agent aura la certitude que son partenaire ne prend pas le tour (y proche de -1),

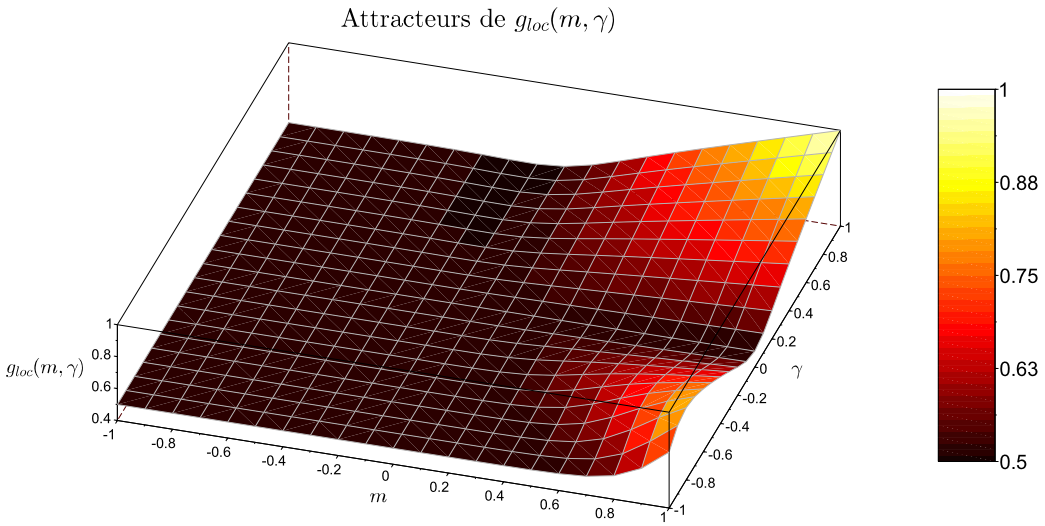


FIGURE 7.3 – Attracteurs de $g_{loc}(m,\gamma)$.

et plus sa motivation à laisser la parole sera grande (m proche de 1), plus il orientera son regard vers son partenaire. Nous définissons ainsi g_{loc} suivant l'équation 7.2. La figure 7.3 illustre l'espace d'état défini par g_{loc}.

$$\begin{aligned} g_{loc} &= 0.5 \\ &+0.5 \times \gamma \times m \times S(\gamma) \times S(m) \\ &-0.5 \times \gamma \times m \times S(\gamma + m) \times S(-\gamma) \times S(m) \end{aligned} \tag{7.2}$$

7.1.2 Actions de l'auditeur

De manière analogue au contrôle de la prosodie par le locuteur courant, l'auditeur courant possède le même espace des attracteurs pour la variation de volume sonore et de hauteur de voix. La valeur par défaut des signaux de hauteur de voix et de volume sonore est de 0. Lorsque la motivation à parler $m > 0$ et $\gamma > 0$, l'agent augmente son niveau sonore et sa hauteur de voix pour signifier sa prise de tour. La prise de tour a lieu ici sans conflit puisque l'agent a un niveau de certitude lui indiquant que le locuteur courant est en train de lui laisser le tour. L'agent module néanmoins la valeur vers laquelle converge son niveau sonore et sa hauteur de voix selon son niveau de certitude. Lorsque l'agent a peu de certitude concernant l'abandon de tour du locuteur courant, il augmente peu son niveau sonore et sa hauteur de voix. Plus γ est positif et grand plus l'agent augmente son niveau sonore avec une valeur proche de 0.5. Lorsque $\gamma < 0$, plus l'agent a une motivation élevée à prendre le tour, plus il cherche à prendre le tour sans attendre d'être plus certain (γ supérieur à 0) que le locuteur courant laisse le tour. Lorsque l'agent a une motivation proche de 1 il augmente son niveau sonore et sa hauteur de voix même si γ est proche de -1. Au contraire lorsqu'il a une motivation faible de parler, il attend d'être plus incertain

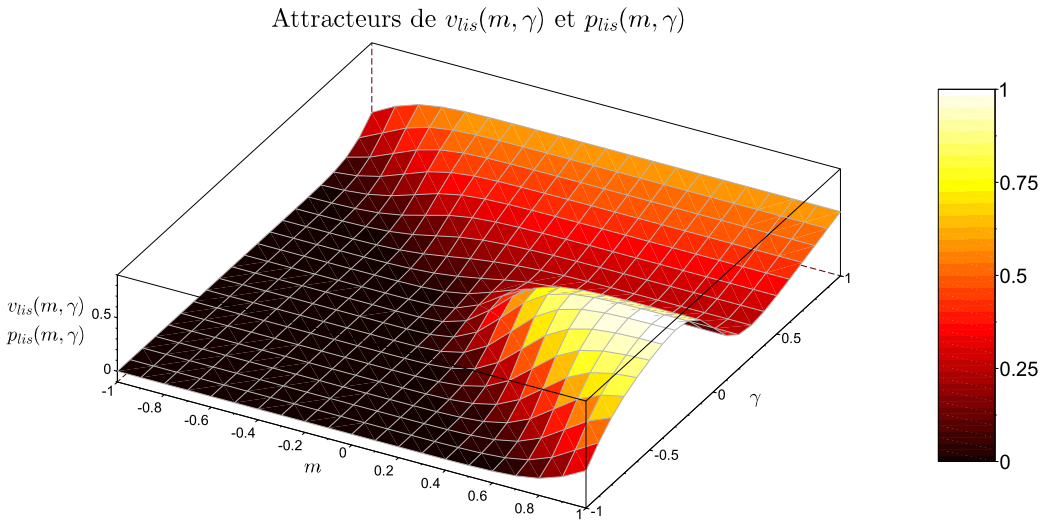


FIGURE 7.4 – Attracteurs de $v_{lis}(m,\gamma)$ et $p_{lis}(m,\gamma)$.

concernant le comportement du locuteur courant pour prendre le tour (γ proche de 0). Lorsque $m < 0$, l'agent peut se mettre à parler si la valeur de confiance γ est suffisamment grande. Nous proposons que plus l'agent a une motivation m proche de -1 (motivation à ne pas parler) plus il prend la parole lorsque le niveau de certitude est élevé. À l'inverse, lorsque l'agent a une motivation $m < 0$ et proche de 0, il commence à augmenter ses signaux pour des valeurs γ proches de 0.

Nous proposons l'équation $v_{lis}(m,\gamma) = p_{lis}(m,\gamma) = \gamma \times 0.5 \times S(\gamma + m)$ pour rendre compte de tous les principes énoncés dans le paragraphe ci-dessus. En situation de conflit ($y < 0$ et $m > 0$), un auditeur ne fait pas que prendre la parole, mais il la prend avec une valeur de volume sonore et de hauteur de voix élevée. La valeur de ces signaux converge donc vers 1. Nous ajoutons donc 0.5 à la valeur de l'attracteur déterminée par cette fonction. Les fonctions v_{lis} et p_{lis} sont alors définies par l'équation 7.3. L'espace des attracteurs résultant est illustré sur la figure 7.4.

$$v_{lis}(m,\gamma) = p_{lis}(m,\gamma) = 0.5 \times S(\gamma + m) \tag{7.3}$$

Pour le contrôle du regard de l'agent, nous proposons une valeur par défaut de 1 illustrant un auditeur regardant le locuteur courant. Lorsqu'il s'apprête à prendre la parole, il détourne le regard. Le taux de regard vers le partenaire de l'agent converge alors vers 0 dans notre modèle. Peu importe le signe de m, la valeur de la direction de son regard dépend de m et γ de la même manière que proposé pour le contrôle de la prosodie : lorsque le niveau de certitude γ est supérieur à m l'agent détourne son regard, lorsqu'il est inférieur, l'agent continue à regarder le locuteur courant. La fonction résultant de ce principe est définie par l'équation 7.4 et illustrée sur la figure 7.5.

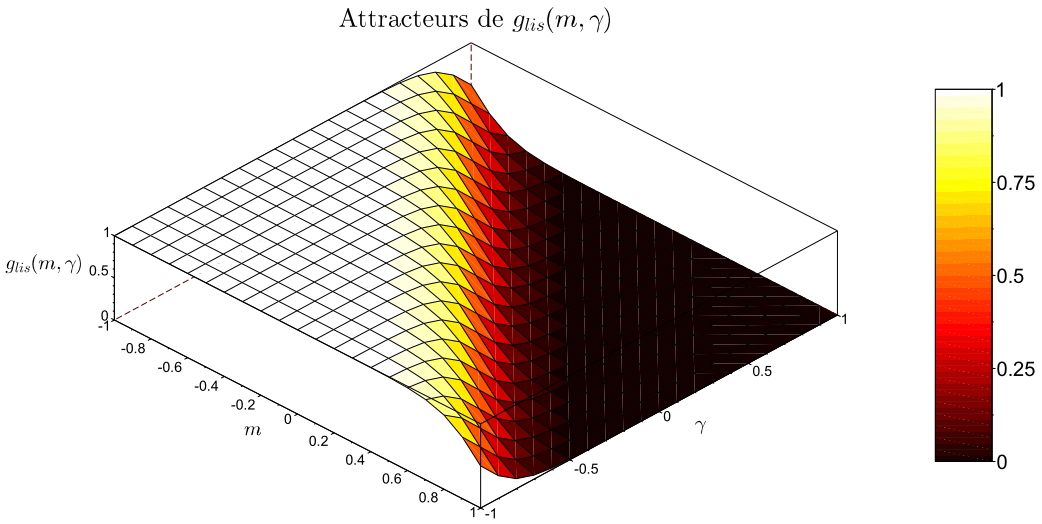


FIGURE 7.5 – Attracteurs de $g_{lis}(m,\gamma)$.

$$g_{lis}(m,\gamma) = 1.0 - S(\gamma + m) \tag{7.4}$$

7.2 Émergence du comportement

Nous illustrons dans cette section l'hypothèse 6 d'émergence du comportement des agents introduit dans le chapitre 6. Nous avons déterminé les fonctions définissant les attracteurs des signaux de sorte d'illustrer l'influence à la fois de la motivation et du niveau sonore sur le contrôle des actions de l'agent. Les attracteurs des actions sont ainsi définis non seulement par la motivation à parler de l'agent mais aussi par le niveau de certitude. Même si l'agent a une motivation constante tout au long de l'interaction, la valeur des attracteurs variera toujours selon le niveau de certitude de l'agent. La variation des signaux de l'agent est donc bien déterminée non seulement par les buts de l'agent mais aussi directement par les variations de signaux produites par l'autre participant.

Nous illustrons ce principe par un exemple issu d'une simulation entre un agent ayant le rôle de locuteur et un agent ayant le rôle d'auditeur. L'agent locuteur a une motivation $m_{loc} = -1$ et l'agent auditeur a une motivation à prendre la parole $m_{lis} = 1$. Comme illustré sur la figure 7.6, lorsque nous simulons ce cas de figure nous obtenons un conflit entre les deux agents, l'auditeur finissant par prendre le tour.

Ce premier scénario montre la dépendance continue dans la production des signaux de chaque agent envers les signaux produits par son partenaire. Le locuteur courant garde au début de cette simulation son niveau sonore à 0.5. À mesure que

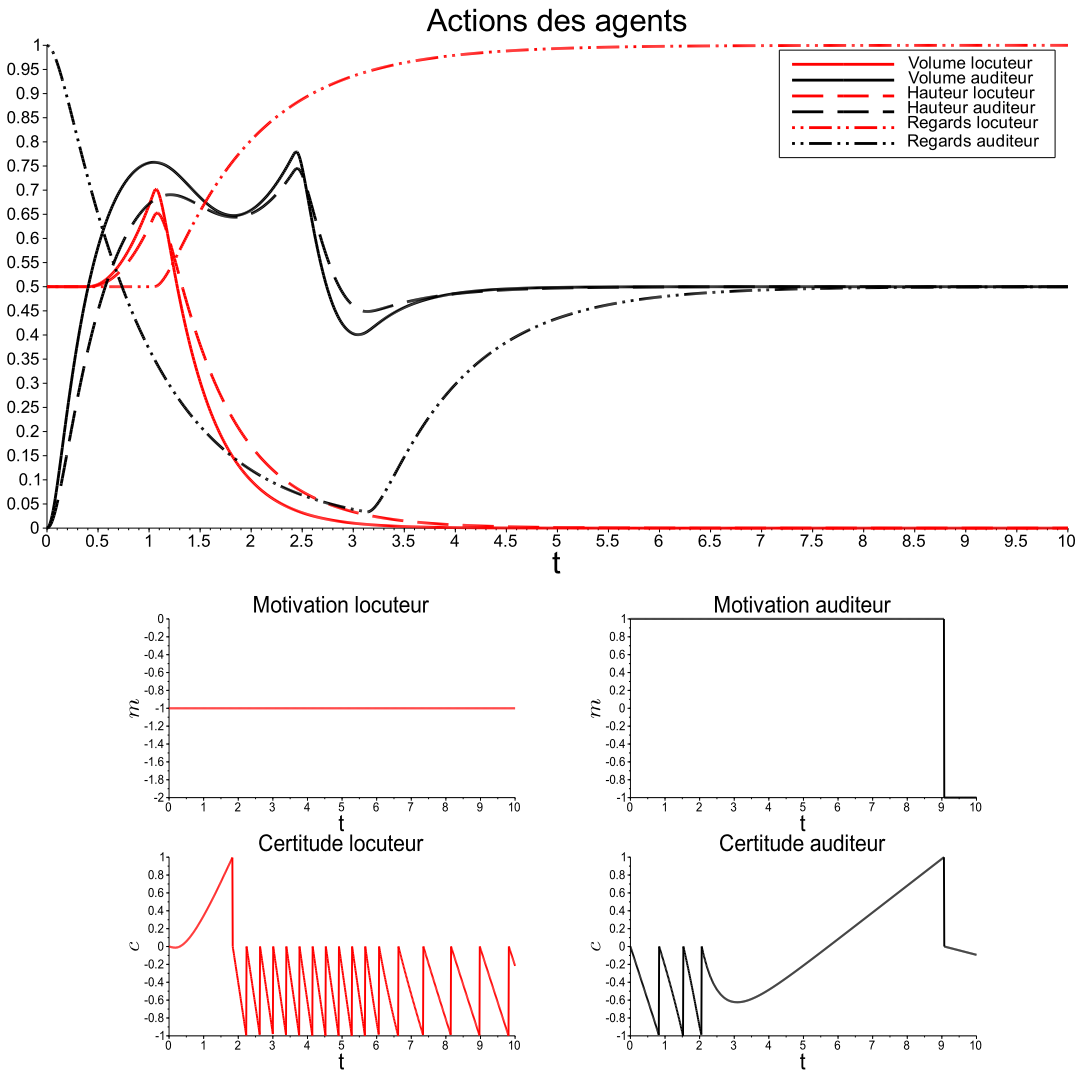


FIGURE 7.6 – Scénario illustrant un conflit entre le locuteur courant et l'auditeur courant (m_{loc} à -1.0 et m_{lis} à 1.0)

son niveau de certitude sur la prise de tour de l'auditeur courant augmente, dû à l'augmentation du volume sonore et de la hauteur de voix de ce dernier, il finit par augmenter son volume sonore pour empêcher l'auditeur courant de prendre la parole. Néanmoins, l'auditeur courant ne diminuant pas ses variables prosodiques, et le niveau de certitude atteignant la valeur seuil 1, l'agent change de rôle, comme défini dans le chapitre 6. Il finit donc par céder et arrêter de parler. Le même constat peut être fait pour les variations de signaux de l'auditeur courant. Il prend d'abord la parole en tenant peu compte des signaux de l'autre agent. On observe ensuite une baisse des valeurs prosodiques de l'agent provenant du fait que la certitude de l'agent reste négative. Puis le volume sonore et la hauteur de voix se stabilisent à la valeur 0.5 à mesure que l'agent devient plus certain sur le fait que le locuteur courant lui cède le tour. Le fait que l'agent locuteur ne cherche pas à reprendre la parole provient du fait que la simulation est effectuée pour un échange de tour, une fois qu'un agent a changé de rôle, nous forçons sa valeur de motivation à -1.0 de sorte qu'il ne cherche pas à changer de rôle ensuite. La même remarque peut être faite pour l'auditeur, lorsqu'il change de rôle, on observe une modification de la valeur de la motivation de 1 à -1.

Si l'on garde la même motivation à parler pour le locuteur courant mais que l'on modifie la valeur de la motivation à parler de l'auditeur nous obtenons un comportement différent tel que montré par la figure 7.7. Ici le locuteur courant réussit à garder la parole malgré les tentatives de prise de parole de l'auditeur courant. L'auditeur courant ayant une motivation plus faible à parler, la valeur de ses signaux prosodiques est moins élevée que précédemment et l'agent se ravise plus vite lorsqu'il n'observe pas d'indices indiquant que le locuteur courant est en train de lui laisser le tour. Le niveau de certitude du locuteur courant augmente puis baisse sans atteindre la valeur 1 dû au fait que les signaux de prise de tour sont moins forts et plus brefs. L'agent garde ainsi la parole. La répétition observée des tentatives de prises de tour provient de la réinitialisation à 0 de la valeur de certitude de l'auditeur courant. Ce dernier redevient incertain sur la nature du comportement du locuteur et augmente de nouveau la valeur de ses signaux.

La modification de la valeur de motivation du locuteur courant impacte de la même manière les comportements des deux participants. Si l'on modifie la motivation à $m = 1$ le locuteur courant laisse le tour à l'auditeur courant résultant en une prise de tour observée sur la figure 7.8.

Ces simulations vérifient l'hypothèse 6, le comportement résultant de chaque participant n'est pas uniquement déterminé par la motivation de l'agent mais aussi indirectement par la motivation de son partenaire. En ce sens un couplage sensori-moteur s'instaure entre les participants, une variation des actions d'un participant impacte directement la variation des actions de l'autre participant. Puisque les participant s'influencent mutuellement, on ne peut déterminer à l'avance le comportement

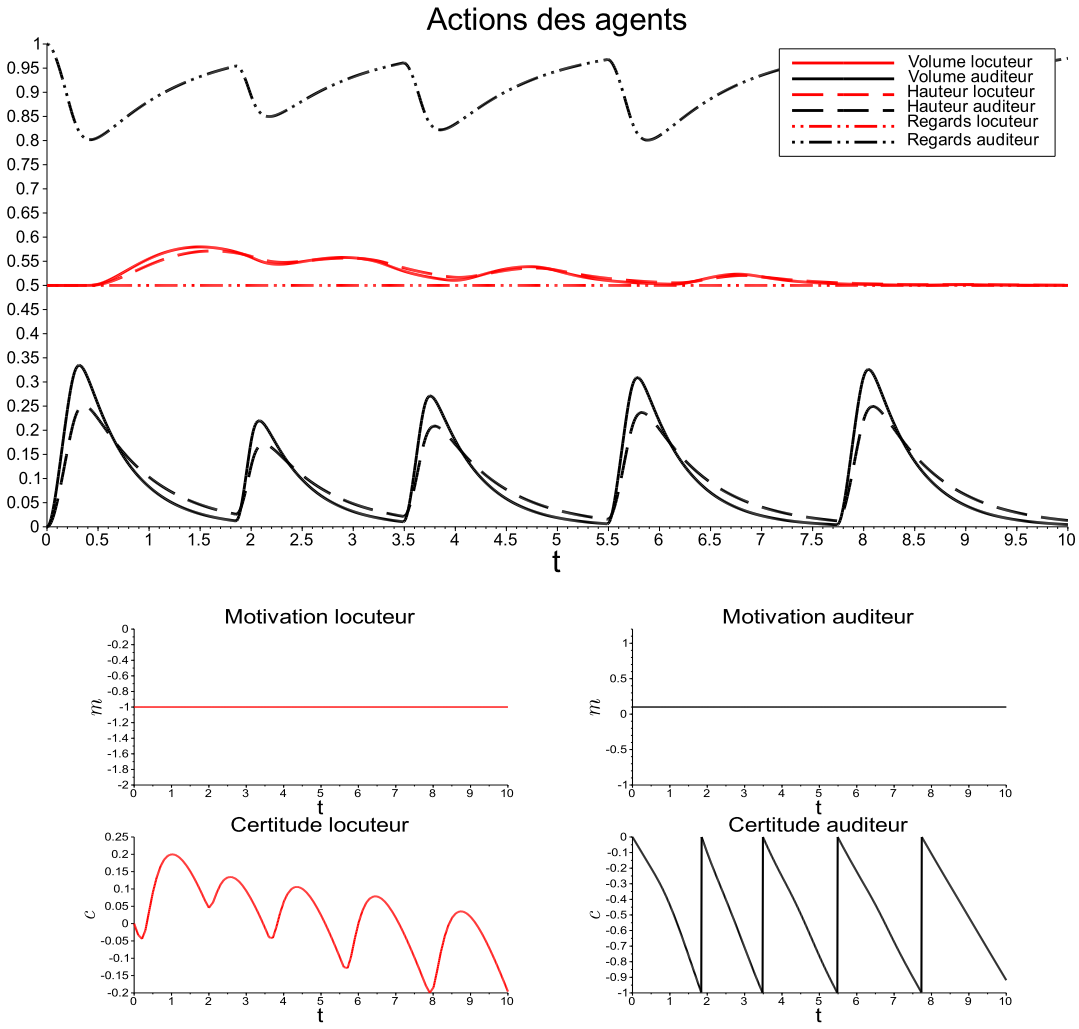


FIGURE 7.7 – Illustration d'un scénario de conflit entre les deux participants où le locuteur courant a une motivation $m_{loc} = -1.0$ et l'auditeur courant a une motivation $m_{lis} = 0.1$

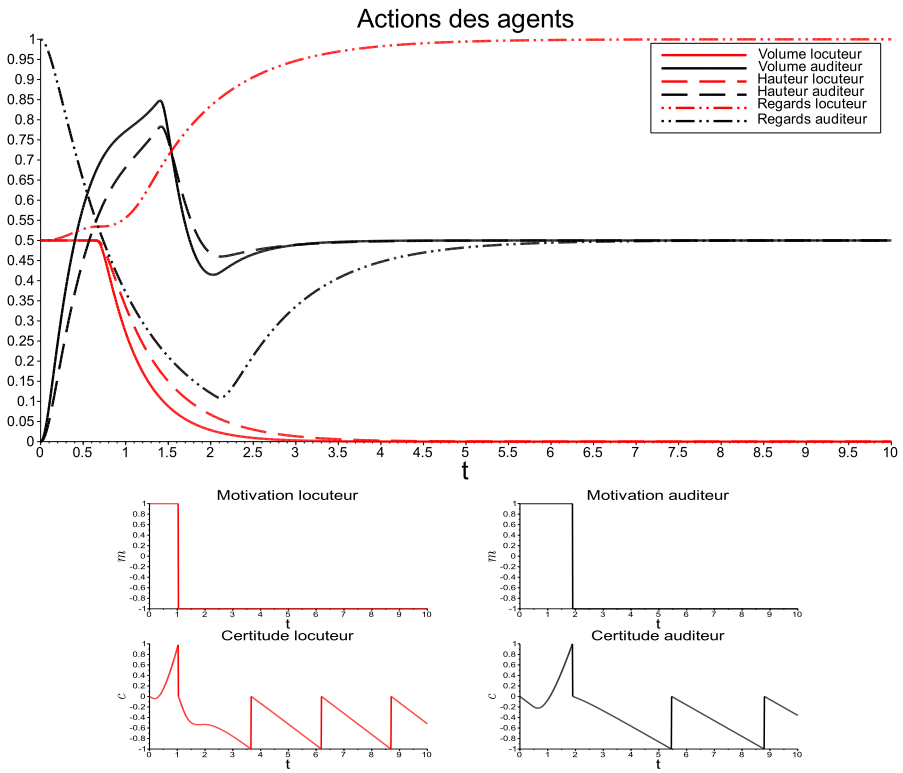


FIGURE 7.8 – Illustration d'une transition « fluide » avec un léger overlap entre les deux participants où le locuteur courant a une motivation $m_{loc} = 1.0$ et l'auditeur courant a une motivation $m_{lis} = 1.0$

du participant, ce dernier est émergent de l'interaction.

7.3 Adaptabilité de l'agent

Nous avons montré le couplage existant entre les deux agents dans la section précédente. Si nous changeons la dynamique de la production des signaux produits par un agent, nous changeons la dynamique de la production des signaux de l'autre agent. Cela implique que non seulement la valeur de motivation peut impacter le comportement de l'autre agent mais également que toute modification dans les paramètres du modèle de l'autre agent devrait modifier l'évolution des actions de l'agent. Nous mesurons cet effet en modifiant les paramètres d'amortissement b, de raideur k_g et les coefficients d'accumulation d'un participant pour une simulation de conflit de parole entre les deux participants. Nous reprenons une valeur de motivation de $m_{lis} = 0.1$ pour l'agent auditeur et $m_{loc} = -1.0$ pour l'agent locuteur donnant les valeurs de la figure 7.7. Prenons comme valeurs d'amortissement $b_v = 20.0$, $b_p = 14.0$, $b_g = 10.0$ et de raideur $k_g^v = 40.0$, $k_g^p = 20.0$ et $k_g^{gd} = 10.0$, où b_v et k_g^v représentent l'amortissement et la raideur du volume sonore, b_p et k_g^p représentent l'amortissement et la raideur de la hauteur de voix et b_g et k_g^{gd} l'amortissement et la raideur de la direction du regard. Si nous implémentons un agent auditeur tel que les paramètres d'amortissement soient égaux à la valeur d'amortissement de l'agent

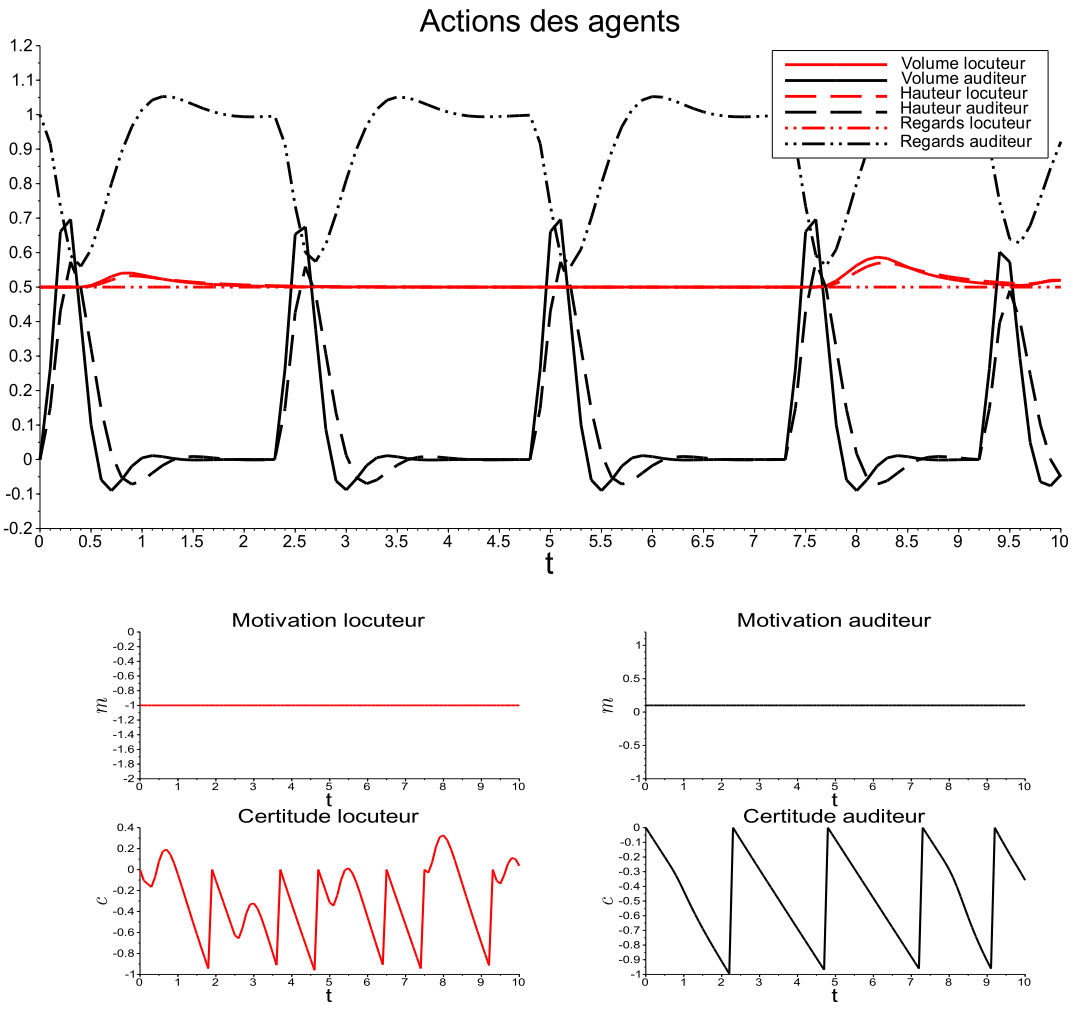


FIGURE 7.9 – Évolution des actions des deux agents au cours du temps avec un agent auditeur ayant des valeurs d'amortissement $b_{slis} = \frac{b_{sloc}}{2}$ et de raideur $k_g^{slis} = 2 \times k_g^{sloc}$ avec *slis* et *sloc* un signal respectivement de l'auditeur et du locuteur évoluant au cours de l'interaction

locuteur, nous obtenons le résultat présenté sur la figure 7.7. Prenons maintenant des valeurs d'amortissement pour l'agent auditeur telles que celles-ci soit égales à la moitié des valeurs d'amortissement de l'agent locuteur pour chaque signal et un paramètre de raideur deux fois plus grand que la raideur de l'agent locuteur. Nous modélisons ainsi un agent auditeur plus réactif que l'agent locuteur : il possède une inertie deux fois moins grande à suivre la modification d'un attracteur et converge deux fois plus rapidement vers l'attracteur. Nous obtenons alors les résultats illustrés sur la figure 7.9. La modification de l'amortissement et de la raideur impacte ici la force avec laquelle l'agent cherche à prendre le tour. Les tentatives de prise de parole sont plus courtes, conséquence de la diminution de l'amortissement rendant l'auditeur plus réactif lorsque l'attracteur diminue vers la valeur 0 pour le volume sonore et la hauteur de voix et augmente vers 1 pour la direction du regard. La force de la tentative de prise de tour est donc plus grande. Afin de conserver la parole, l'agent accentue sa variation de signaux pour signifier son désir de garder la parole.

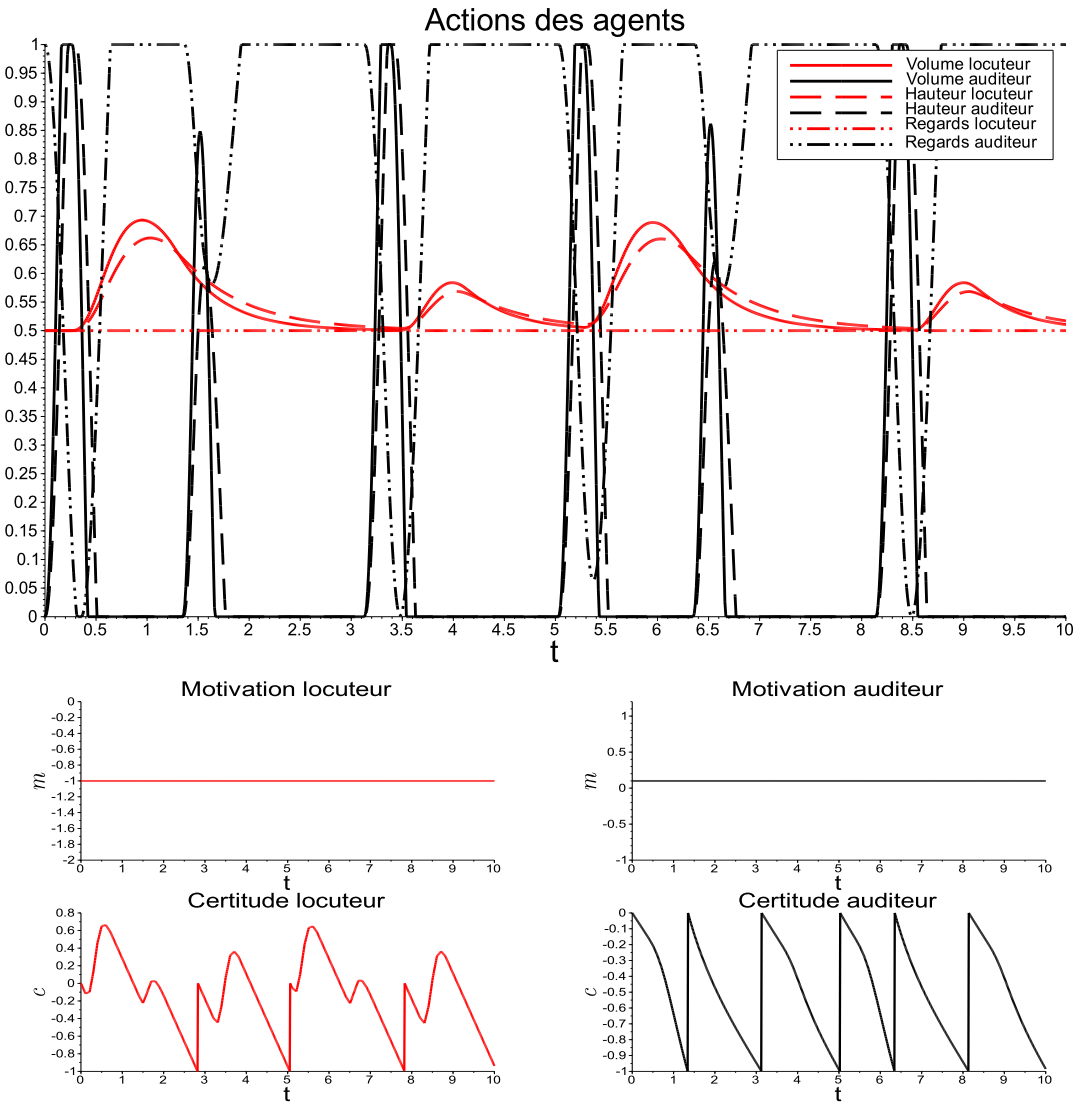


FIGURE 7.10 – Illustration d'une simulation avec $b_{slis} = \frac{b_{sloc}}{4}$ et $k_g^{slis} = 4 \times k_g^{sloc}$

Si l'on diminue la valeur de l'amortissement tel que $b_{slis} = \frac{b_{sloc}}{4}$ et que l'on augmente la valeur de raideur tel que $k_g^{slis} = 4 \times k_g^{sloc}$, nous constatons que l'agent locuteur parvient toujours à conserver son rôle mais accentue encore la variation de la hauteur de voix et du volume sonore pour y parvenir, tel que montré sur la figure 7.10. Modifier l'amortissement et la raideur d'un facteur quatre dégrade fortement la production de signaux de l'agent : nous ne nous trouvons plus en régime apériodique et la production de signal de l'agent oscille autour des valeurs 0 et 1, nous sommes donc à la limite de notre modèle. Pour éviter d'avoir des valeurs de signaux négatives ou plus grandes que 1, nous forçons les valeurs négatives des signaux de l'agent auditeur à 0 et les valeurs plus grandes que 1 à 1 résultant en une cassure dans la variation des signaux des agents observée sur la figure.

La même adaptation se vérifie pour un scénario où le locuteur a une motivation à parler $m_{loc} = 1.0$, l'auditeur ayant une motivation à parler $m_{lis} = 0.3$. L'agent laisse la parole plus ou moins rapidement selon l'amortissement et la raideur de l'agent

auditeur tel que montré sur la figure 7.11.

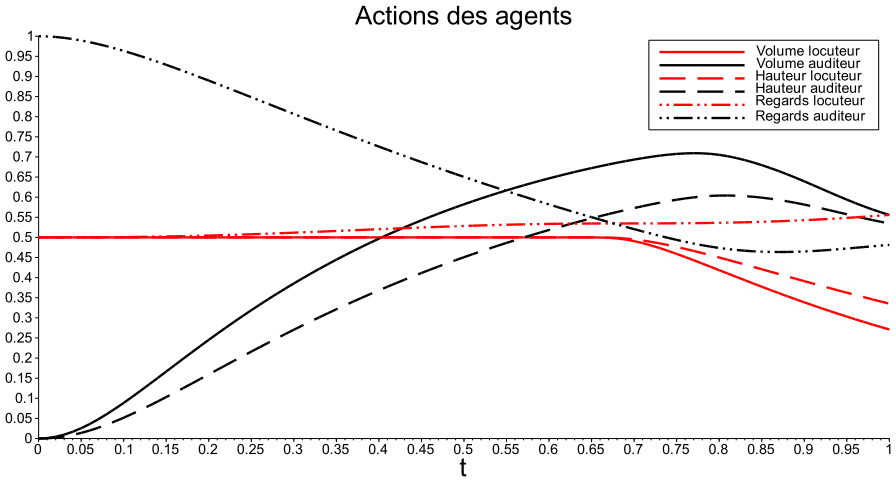
Observons maintenant l'effet d'un agent percevant plus ou moins rapidement la variation des actions. Nous modifions les capacités de perception de l'agent en multipliant ou divisant la valeur d'accumulation résultante. Nous obtenons alors un agent dont le degré de certitude varie plus ou moins rapidement. Prenons le même scénario de conflit que celui présenté ci-dessus. Nous divisons par deux la valeur résultante de la fonction d'accumulation de l'agent auditeur tel que l'équation de perception du comportement peut être réduit à l'équation 7.5.

$$d\gamma = \frac{\alpha(t)}{2} \times dt + \sigma \times dW \tag{7.5}$$

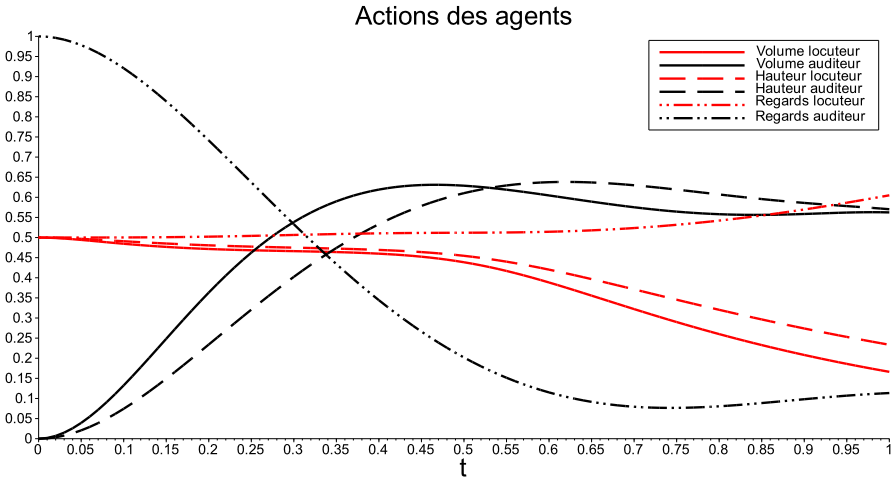
Nous obtenons alors la variation d'action montrée sur la figure 7.12. L'agent ayant une valeur d'accumulation divisée par deux, son niveau de certitude décroit plus lentement. L'auditeur reste plus longtemps incertain sur le comportement du locuteur. Il augmentera plus longtemps ses signaux prosodiques avant de se raviser, ce qui résulte en une valeur de ces signaux plus grande que précédemment. L'auditeur augmentant plus ses signaux de prise de tour, le locuteur a une valeur de certitude plus grande, ce qui résulte en une accentuation de la variation de ses signaux pour signifier qu'il garde le tour.

Lorsque nous diminuons par quatre la valeur d'accumulation nous obtenons les résultats montrés sur la figure 7.13. Dans ce scénario, la valeur de certitude du locuteur courant atteint le seuil positif avant que l'auditeur soit plus certain sur la volonté du locuteur courant de garder le tour. Le locuteur courant devient auditeur. L'auditeur ne repère pas tout de suite la fin de parole du locuteur, il commence à se raviser puis augmente la valeur de ses signaux lorsqu'il repère enfin la fin de parole du locuteur courant.

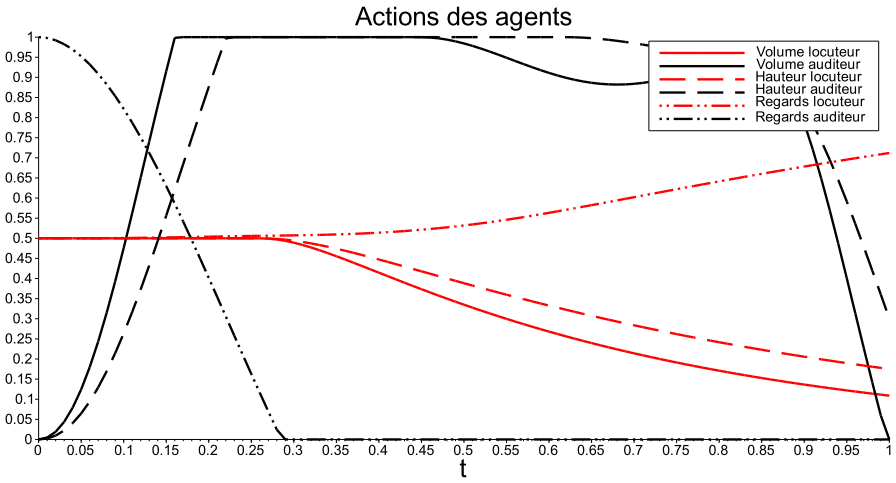
Nous avons montré la capacité des agents à s'adapter à différents types de partenaires conversationnels en modifiant les variables d'amortissement et de raideur d'un côté et d'accumulation de l'autre. Ces derniers sont plus ou moins réactifs aux variations de signaux de l'autre et à la variation de leurs attracteurs selon leur coefficient d'accumulation, leur amortissement et leur raideur. Afin de garantir la satisfaction de leur but, les agents sont capables de compenser des modifications dans la variation de leurs signaux par une accentuation de leurs propres signaux. La capacité à adapter la production de leurs signaux ne peut se vérifier que si les participants sont couplés dans la production de leurs signaux. Si la variation de signal provenant d'une modification de b_s et k_g^s ne se répercutait pas directement sur la propre production de signaux du locuteur, ce dernier n'aurait pas augmenté la valeur de ses signaux, qui en retour n'aurait pas fait décroitre plus rapidement le niveau de certitude de l'auditeur courant et poussé ce dernier à se raviser plus rapidement. Cette capacité d'adaptation n'a pas été énoncée dans le modèle, c'est une capacité émergente de l'interaction entre les participants.



(a) Échange de tour avec $b_{slis} = b_{sloc}$ et $k_g^{slis} = k_g^{sloc}$



(b) Échange de tour avec $b_{slis} = \frac{b_{sloc}}{2}$ et $k_g^{slis} = 2 \times k_g^{sloc}$



(c) Échange de tour avec $b_{slis} = \frac{b_{sloc}}{4}$ et $k_g^{slis} = 4 \times k_g^{sloc}$

FIGURE 7.11 – Illustrations d'un scénario d'échange de tour pour des valeurs d'amortissement et de raideur différentes

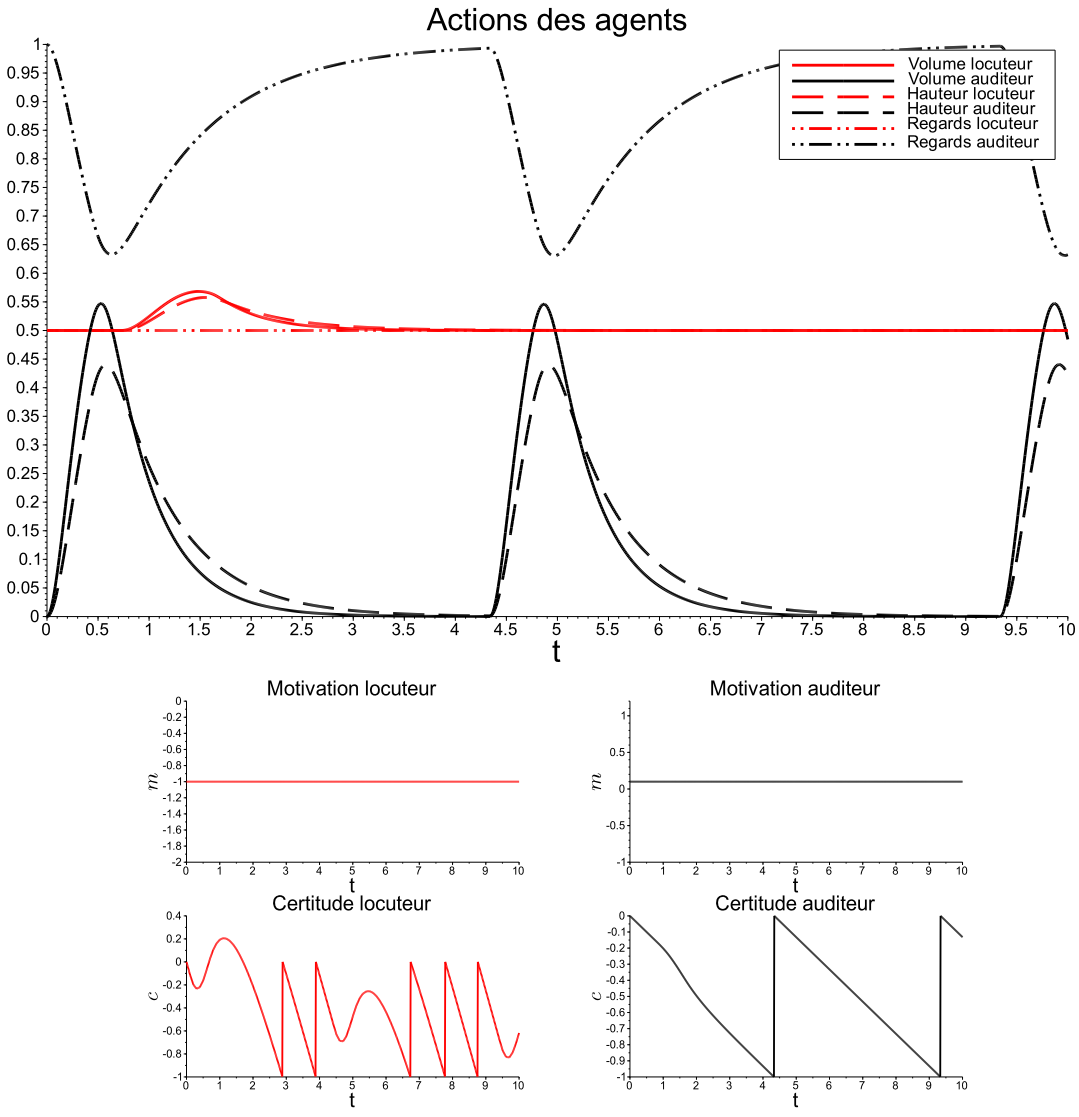


FIGURE 7.12 – Simulation d'un scénario de conflit avec la valeur d'accumulation de l'auditeur divisée par deux ($m_{loc} = -1.0$ et $m_{lis} = 0.1$).

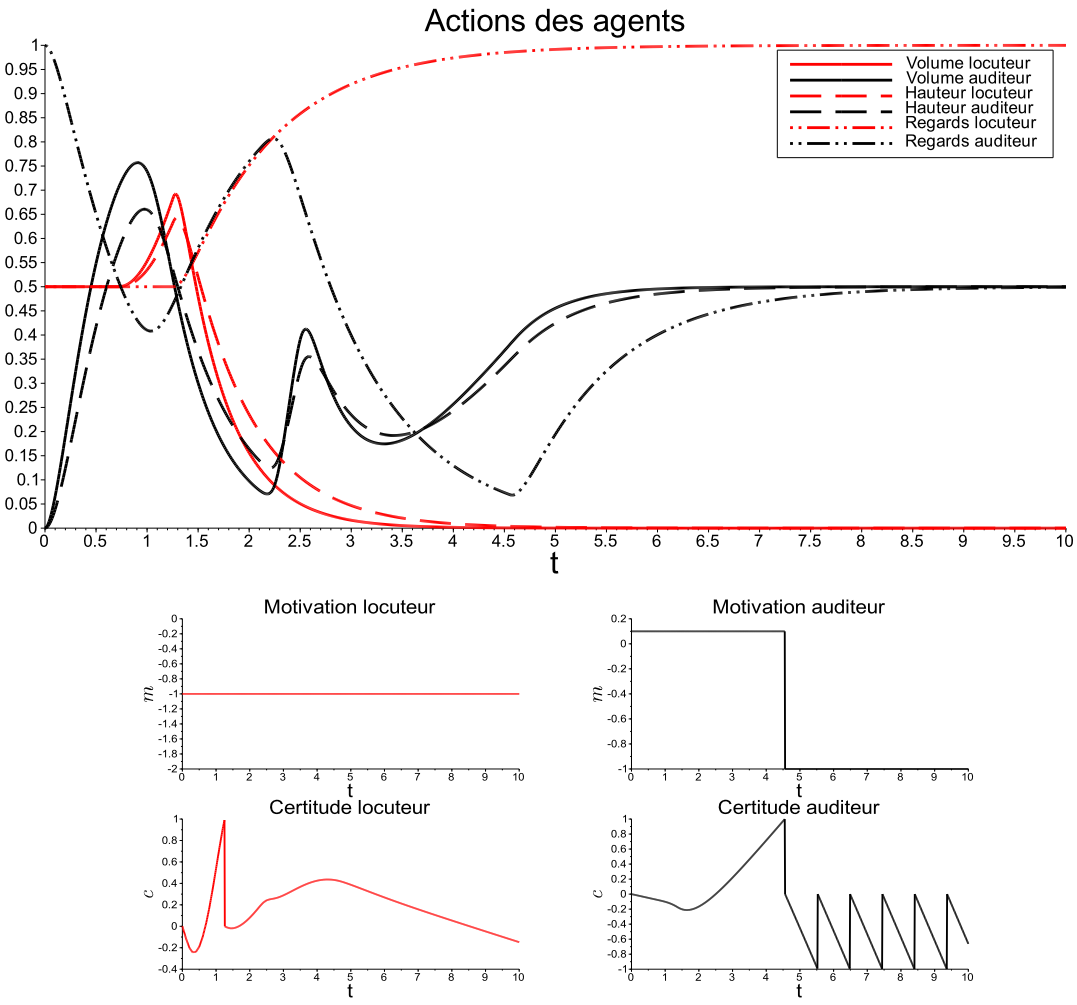


FIGURE 7.13 – Simulation d'un scénario de conflit avec la valeur d'accumulation de l'auditeur divisée par quatre ($m_{loc} = -1.0$ et $m_{lis} = 0.1$).

7.4 Robustesse de la simulation

Ikegami et Iizuka (2007) mettent en avant la capacité des participants couplés à compenser par leur interaction des fluctuations dans l'interprétation des signaux du partenaires. Comme rapporté par De Loor *et al.* (2009), c'est même une caractéristique des systèmes couplés. Ce couplage entre les participants offre une robustesse à d'éventuelles modifications de l'environnement : si les participants n'ont plus accès à certains signaux, ou si l'information apparait bruitée, ces derniers s'adapteront avec peu d'efforts pour permettre la poursuite de la conversation. Dans le cadre d'une gestion du tour de parole entre un utilisateur et un agent, la capacité d'un agent à garantir cette robustesse est un atout majeur pour la conduite de l'interaction. Nous montrons dans cette section la capacité des agents à garantir la coordination des échanges de parole dans différents types d'environnement : des environnements où les participants n'ont accès qu'à un sous-ensemble de signaux ou des environnements où les signaux produits par un participant apparaissent bruités. L'ajustement des participants est assuré sans faire appel à d'autres équations que celles définies dans la section 7.1.

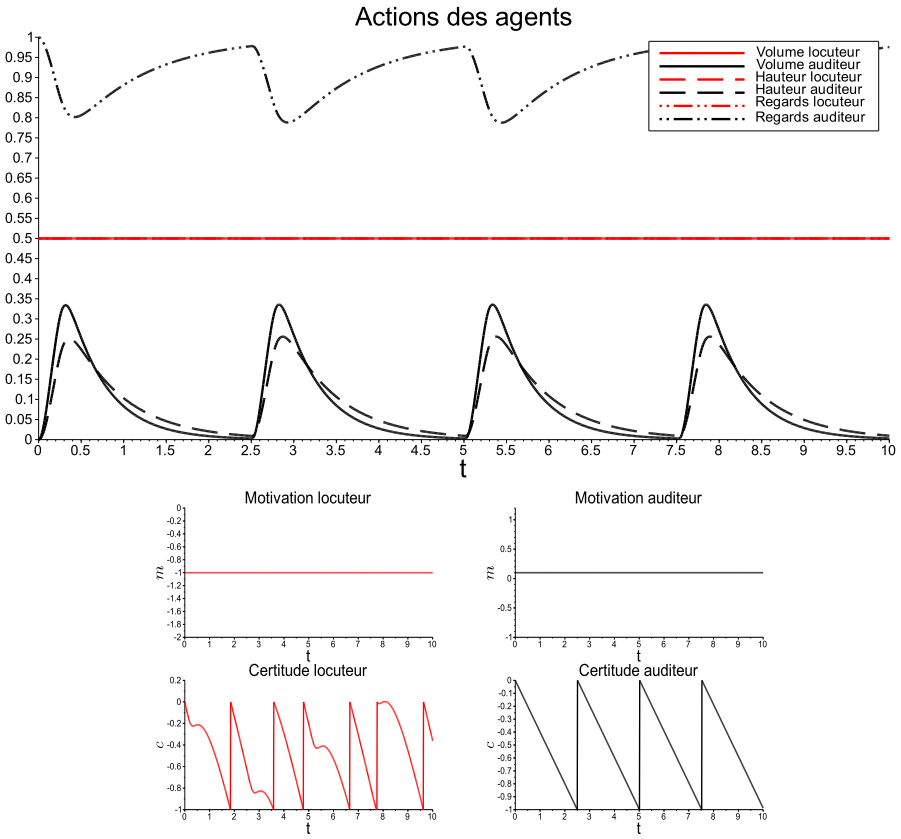
7.4.1 Robustesse à l'absence de signaux

Nous montrons tout d'abord la capacité des participants à se coordonner lorsqu'ils n'ont accès qu'à un sous-ensemble de signaux. Le fait que les participants aient accès à un sous-ensemble de signaux n'implique pas que les participants ne produisent qu'un sous-ensemble de signaux. Pour chaque simulation montrée dans la suite, les participants produisent la totalité des signaux mais ne perçoivent qu'un sous-ensemble de signaux, la valeur d'accumulation partielle pour les signaux non interprétées étant forcée à 0.

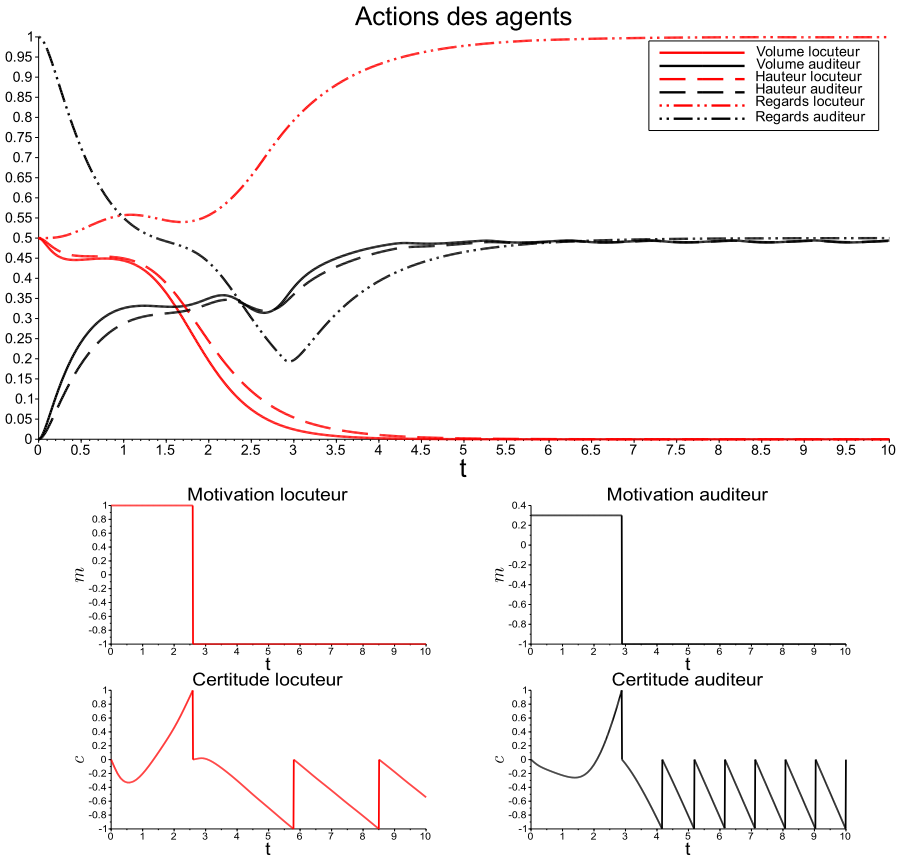
Nous analysons tout d'abord le comportement des agents lorsqu'ils n'ont pas accès aux informations sur la direction du regard de l'autre participant. Cette capacité est montrée sur deux scénarios, le même scénario de conflit que présenté dans les sections précédentes avec $m_{loc} = -1.0$ et $m_{lis} = 0.1$ et un scénario d'échange de tour avec $m_{loc} = 1.0$ et $m_{lis} = 0.3$.

La figure 7.14 illustre la production des signaux des deux participants pour ces deux scénarios lorsque les participants interprètent tous les signaux produits par l'autre participant.

L'évolution des signaux des deux agents lorsque l'information du regard est retirée est montrée sur la figure 7.15. On observe une augmentation des signaux prosodiques du locuteur que l'on n'observait pas lorsque les agents interprétaient la direction du regard. Cela est dû à l'influence conjointe de l'augmentation du volume sonore, de la hauteur de voix de l'auditeur et de la valeur d'accumulation liée à l'absence de la direction du regard. En ce qui concerne l'augmentation de la valeur

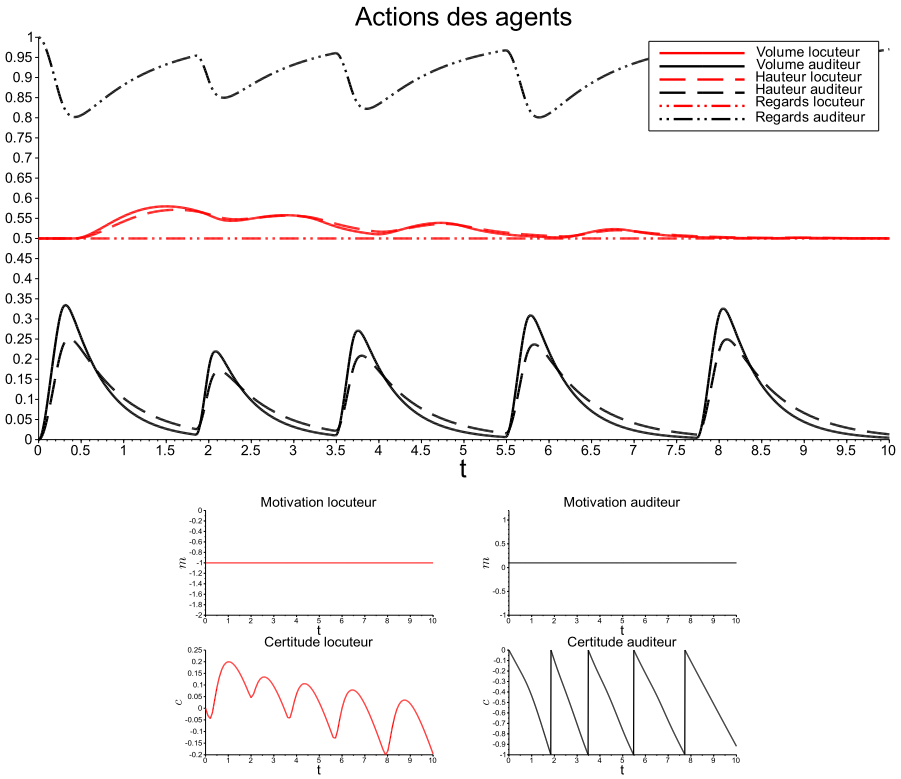


(a) Scénario de conflit où les participants ont accès à tous les signaux produits par leur partenaire. Dans ce scénario $m_{loc} = -1.0$ et $m_{lis} = 0.1$.

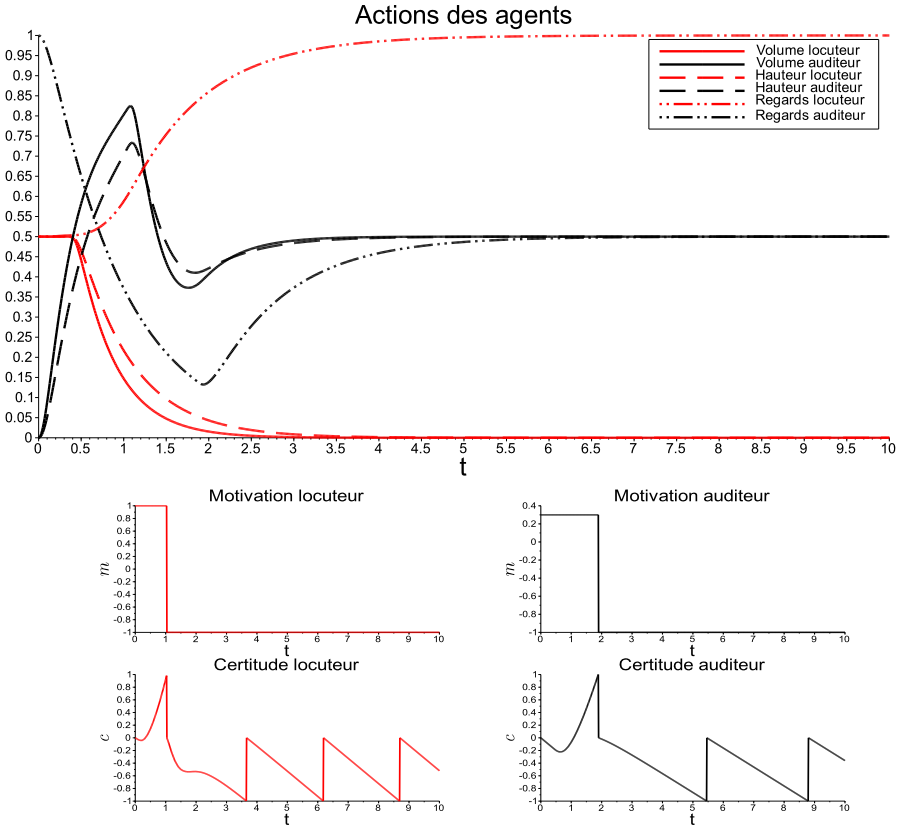


(b) Scénario d'échange de tours où les participants ont accès à tous les signaux produits par leur partenaire. Dans ce scénario $m_{loc} = 1.0$ et $m_{lis} = 0.3$

FIGURE 7.14 – Exemple de deux scénarios où les agents interprètent tous les signaux à leur disposition.



(a) Scénario de conflit où les participants n'interprètent pas la variation de la direction du regard de l'autre participant.



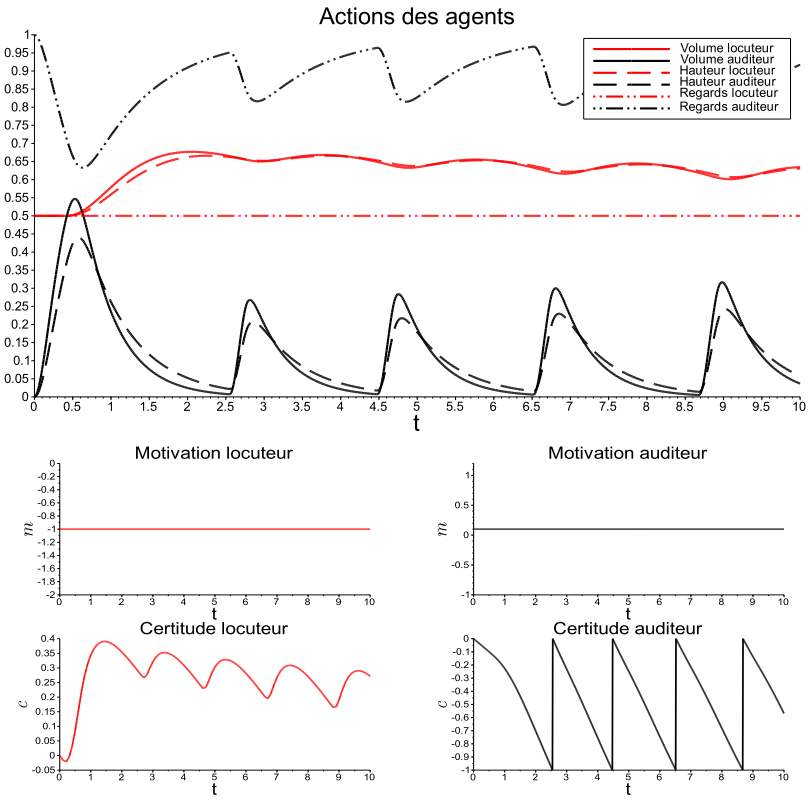
(b) Scénario d'échange de tour où les participants n'interprètent pas la variation de la direction du regard de l'autre participant.

FIGURE 7.15 – Exemple de deux scénarios où les agents n'interprètent pas la variation de la direction du regard de l'autre participant.

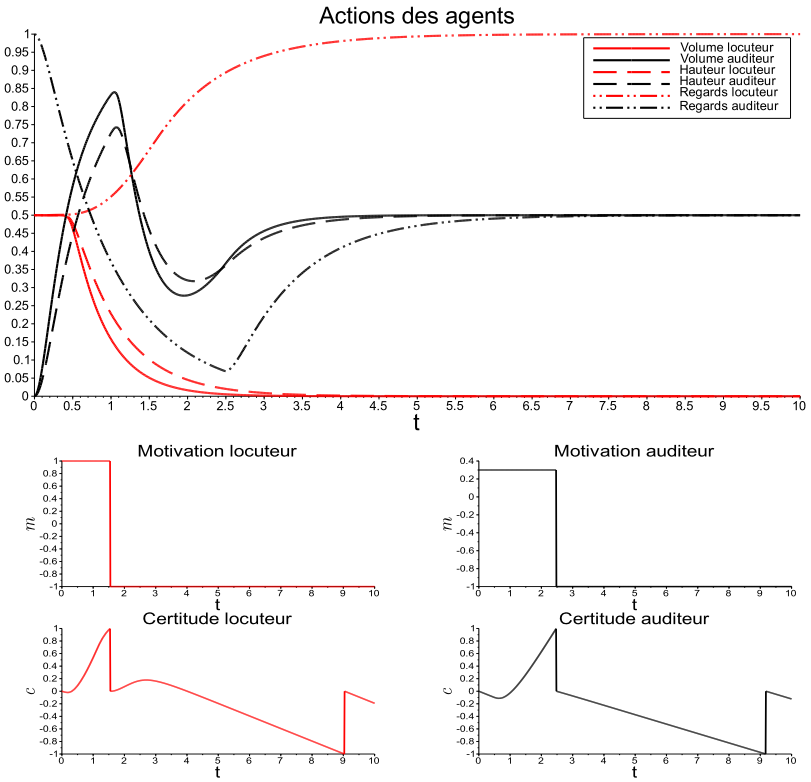
d'accumulation, l'évolution de la direction du regard étant plus lente que l'évolution des deux autres signaux, l'absence d'interprétation du regard a pour effet d'accroitre la variation de la certitude du locuteur. Lorsque le locuteur interprète la direction du regard, la valeur de certitude décroît, même lorsque l'auditeur augmente son volume sonore et sa hauteur de voix en début de simulation. Cela est dû au fait que la direction du regard n'a pas encore assez varié pour résulter en une valeur d'accumulation positive. Lorsque l'on enlève cette information, la valeur d'accumulation est plus grande. Dans le scénario de conflit, enlever la direction du regard a pour effet de rendre positive la valeur d'accumulation. La valeur de certitude augmente alors lorsque l'auditeur essaie de prendre le tour. Une valeur plus grande du volume sonore et de la hauteur de voix de l'auditeur est quant à elle liée à une décroissance plus lente de la certitude de l'agent. L'auditeur reste incertain plus longtemps quant au comportement du locuteur. Cette augmentation du moment d'incertitude, accroît le temps d'augmentation des signaux de niveau sonore et de hauteur de voix vers leur attracteur respectif défini à 1, avant que l'auditeur ne se ravise et ne baisse ses signaux à 0. Aussi, au moment de la bifurcation, la valeur du niveau sonore et de la hauteur de voix a atteint une valeur plus grande que précédemment.

Intéressons-nous à l'échange des tours des participants dans le scénario où les agents ont une motivation à laisser et prendre le tour. Dans notre scénario d'interaction, enlever l'information concernant le regard conduit ici à une transition plus fluide des participants. La certitude de l'auditeur décroit plus rapidement lorsque ce dernier n'interprète pas les variations de regard du participant. En effet, le locuteur tarde plus à fournir des variations de signaux indiquant l'abandon de son tour que dans le scénario avec interprétation de la direction du regard. La raison pour laquelle le locuteur courant a une valeur d'accumulation plus grande lorsqu'il n'a pas accès à l'interprétation du regard est la même que celle mentionnée dans le scénario de conflit ci-dessus. La décroissance plus rapide de l'incertitude de l'auditeur conduit à une augmentation plus forte du volume sonore et de la hauteur de voix. En réponse, la certitude du locuteur courant croit plus rapidement : celui-ci décroit alors ses signaux prosodiques plus rapidement, conduisant à un abandon de tour plus tôt dans la simulation. L'auditeur change alors de rôle plus rapidement.

Intéressons-nous maintenant au comportement des agents lorsqu'ils n'interprètent ni la direction du regard ni la variation de la hauteur de voix. Les résultats des simulations pour les deux scénarios sont représentés sur la figure 7.16. Enlever l'interprétation de la variation de la hauteur de voix décroit la valeur absolue de l'accumulation de chaque participant. Ainsi dans le scénario de conflit montré sur la figure 7.16a, la valeur de certitude de l'auditeur décroit moins rapidement : l'agent reste plus longtemps incertain sur le comportement du locuteur. Il augmente alors ses signaux prosodiques plus fortement que pour les deux scénarios précédents. En réaction, la certitude du locuteur croît plus rapidement, et le locuteur accroit plus



(a) Scénario de conflit où les participants n'interprètent ni la variation de la direction du regard ni celle de la hauteur de voix.



(b) Scénario d'échange de tour où les participants n'interprètent ni la variation de la direction du regard ni celle de la hauteur de voix.

FIGURE 7.16 – Exemple de deux scénarios où les agents n'interprètent ni la variation de la direction du regard de l'autre participant ni la hauteur de voix.

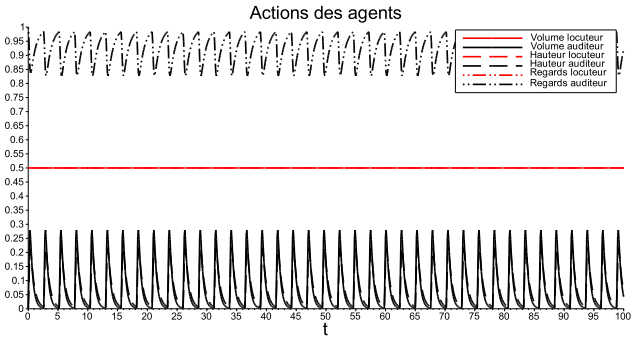
fortement ses signaux prosodiques. L'auditeur se ravise alors plus rapidement. Le fait que le locuteur garde plus longtemps son volume sonore et sa hauteur de voix à une valeur élevée provient de la fréquence des tentatives de prise de tour de l'auditeur, plus grande que précédemment.

7.4.2 Robustesse à un environnement bruité

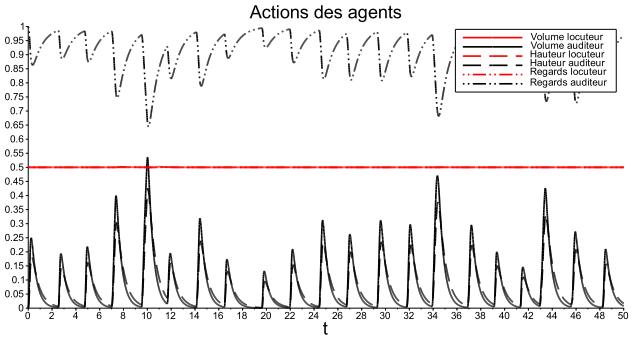
Nous évaluons maintenant la robustesse de la coordination des échanges de parole des agents lorsque l'interprétation du comportement du locuteur est bruitée. La coordination est évaluée par la capacité des agents à garantir une alternance des tours en résolvant rapidement les situations conflictuelles. Le caractère bruité dans l'interprétation des comportements des participants est produit en introduisant un paramètre σ non nul dans l'équation 6.1 du chapitre 6 page 103. Nous illustrons ici cette robustesse en évaluant la durée où les participants parlent en même temps dans le même scénario de conflit que celui étudié jusqu'à présent.

Nous avons augmenté le temps de simulation pour observer la stabilité des comportements au cours du temps.

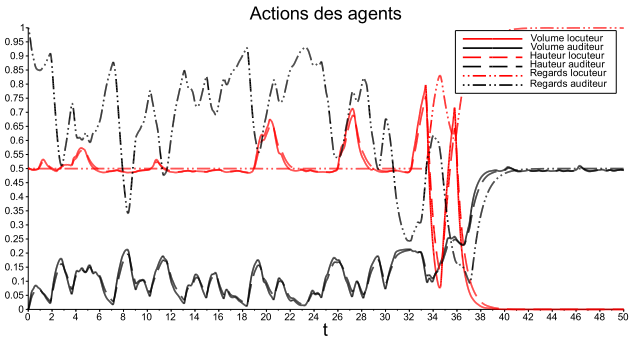
L'augmentation du temps de simulation est une nécessité statistique, nous diminuons ainsi la probabilité d'observer, lors d'une simulation, les comportements les plus courants des participants. Nous présentons différentes simulations illustrées sur la figure 7.17. La figure 7.17a montre le scénario entre deux agents sans fluctuation de leur perception. Si nous fixons $\sigma = 1.0$ nous obtenons un exemple de scénario montré sur la figure 7.17b. On constate que la fluctuation aléatoire influence peu le comportement des participants, son effet sur le processus d'accumulation des deux participants est faible. Nous commençons à observer une instabilité lorsque $\sigma = 4.0$ (figure 7.17c). À cette valeur de σ, les comportements des deux participants sont influencés par les fluctuations aléatoires des processus d'accumulation. Une variation dans le comportement d'un agent liée à une fluctuation du processus de perception est compensée par la variation de l'action de l'autre agent, comme observé au début de la simulation où le locuteur augmente son volume sonore lorsque le processus d'accumulation de l'auditeur amène ce dernier à augmenter de manière significative la valeur de ses signaux prosodiques. Nous constatons néanmoins qu'une fluctuation forte du processus de décision d'un participant peut amener à une transition de tour. Lors de cette transition de tour, les participants sont toujours capables de se coordonner et une nouvelle situation stable apparait alors, avec l'auditeur passé locuteur et inversement. La figure 7.17d montre quant à elle la simulation d'un scénario avec $\sigma = 8.0$. Ici le processus d'accumulation des participants devient fortement dégradé par la fluctuation aléatoire. Les participants ne parviennent plus à compenser cette fluctuation aléatoire et peinent à parvenir à une situation stable.



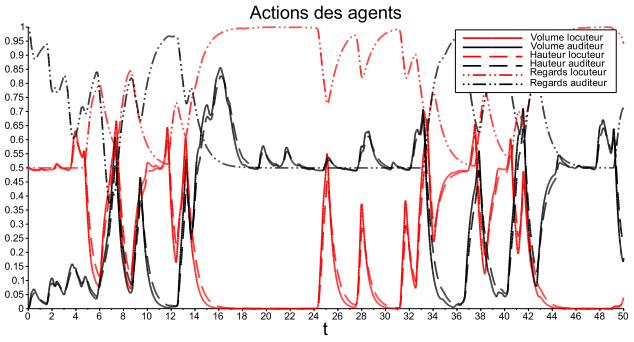
(a) Scénario de conflit sans fluctuation aléatoire dans l'interprétation des signaux.



(b) Scénarios de conflit avec une fluctuation aléatoire $\sigma = 1.0$.



(c) Scénario de conflit avec une fluctuation aléatoire $\sigma = 4.0$.



(d) Scénario de conflit avec une fluctuation aléatoire $\sigma = 8.0$.

FIGURE 7.17 – Scénario de conflit avec différentes valeurs de σ.

7.5 Conclusion

Nous avons analysé dans ce chapitre le comportement du modèle dans le cadre de la simulation d'une interaction agent-agent. En mettant en place nos équations d'accumulation et de contrôle des actions des participants, nous avons créé les conditions de couplage sensorimoteur des participants sans spécifier l'issue de l'interaction pour m_{loc} et m_{lis} donné. Nous avons illustré à partir de ces équations la capacité du modèle à faire émerger les comportements des participants comme prévu par l'hypothèse 6. De plus, en mettant en place ce couplage entre les participants nous avons montré, sans modifier les équations utilisées par notre modèle, la capacité de l'agent à adapter son comportement à différents types de partenaires et sa capacité à s'adapter lorsqu'il n'a accès qu'à un sous-ensemble de signaux et dans un environnement modérément bruité. L'adaptation et la robustesse de l'agent sont une qualité de notre modèle, dans la mesure où dans le cadre d'une interaction agent-utilisateur l'agent n'interagit que rarement tout le temps avec le même participant et peut parfois avoir des données bruitées liées à la qualité de ses capteurs. Le comportement du modèle en interaction temps réel avec l'utilisateur est traité au chapitre 10.

Chapitre 8

Architecture BeAware

Dans les chapitres 6 et 7, nous avons présenté notre modèle théorique de coordination du tour de parole que nous avons appliqué à une interaction simulée entre deux agents théoriques. Pour aller vers l'implémentation d'un agent gérant le tour de parole dans une interaction dialogique avec l'utilisateur nous devons déterminer comment intégrer notre modèle de tour de parole dans un contexte réel de dialogue agent-utilisateur. Cela implique notamment de déterminer quelle architecture informatique est la plus adaptée pour l'implémentation d'un modèle continu de gestion du tour de parole. Pour répondre à ces questions nous nous penchons dans ce chapitre sur la conception d'une architecture informatique d'agent supportant l'implémentation de notre modèle. Nous présentons dans une première partie les pré-requis nécessaires pour l'implémentation du modèle présenté dans les chapitres précédents puis l'architecture BeAware que nous avons créée à partir des architectures ASAP et Ymir. La manière dont le modèle est implémenté dans l'architecture BeAware sera présentée dans le chapitre 10.

8.1 Problématiques d'implémentation du modèle

8.1.1 Perception et action continue

Une des caractéristiques les plus importantes de notre modèle est son caractère continu. La continuité se réfère ici à deux notions différentes, d'une part le caractère continu de la perception de l'agent et d'autre part la continuité dans la production d'action. La perception continue se réfère plus précisément dans notre cas au fait que l'agent interprète des informations de nature continue temporellement (hauteur de voix, volume sonore par exemple) et attribue un degré de certitude continu sur le comportement de son partenaire (prendre le tour, laisser le tour, garder le tour ...). Cela implique d'avoir, à la fois, une architecture capable de traiter des variables numériques continues mais aussi capable de gérer un flux continu de données provenant des capteurs de l'agent.

Le caractère émergent de notre architecture implique que l'agent ne peut connaitre à l'avance quand l'action va finir. En ce sens, aucune limite temporelle n'est définie pour l'action qui est exécutée tant qu'un ordre d'arrêt n'a pas été reçu.

La continuité s'exprime enfin dans le cycle de perception-action dans lequel les agents sont engagés, l'agent agit et perçoit sans alternance entre le processus de perception et d'action. Les deux processus fonctionnent en parallèle plutôt que de manière séquentielle, et une variation dans le processus de perception induit directement une variation dans l'action. Dans le cadre d'une interaction utilisateur-agent, cela implique que l'agent n'attend pas la fin de l'action de l'utilisateur avant de l'interpréter et de commencer à réagir à cette action. Tel que vu dans le chapitre 2, seules les architectures incrémentales d'agent sont capables de disposer de telles capacités.

8.1.2 Génération multimodale d'actions

La génération multimodale d'actions implique la capacité à produire de manière synchronisée plusieurs actions motrices sur différents canaux de communication. Notre modèle de gestion du tour de parole gère une partie de ces contraintes de multimodalité par la synergie existant entre les différentes équations de contrôle de l'action, telle que présentée dans le chapitre 6. Les actions implémentées jusque-là ont été unimodales. Nous envisageons la possibilité d'un contrôle dynamique et continu d'une action multimodale dans notre architecture. Cette action multimodale, abstraite, doit être décomposée en un certain nombre d'actions concrètes générées par les actionneurs de l'agent. Notre architecture d'agent doit avoir non seulement la possibilité de traduire une action en un ensemble de commandes motrices mais aussi de pouvoir modifier dynamiquement une commande en cours de réalisation par l'agent en réponse à des modifications de l'état de l'environnement ou de variables internes à l'agent (motivation à parler par exemple).

8.1.3 Contrôle parallèle de l'action et gestion des ressources corporelles de l'agent

Notre modèle de gestion du tour de parole repose sur plusieurs équations de contrôle de signaux s'exécutant en parallèle, ce qui génère des contraintes de synchronisation et d'accès à des ressources partagées qui doivent être gérées par l'architecture. Dans les scénarios présentés dans les chapitres 6 et 7, le contrôle de la prosodie nécessite que l'agent parle, c'est-à-dire que le volume sonore de l'agent ne soit pas à 0. Il est donc nécessaire de s'assurer que cette contrainte soit bien respectée. Imaginons maintenant que deux actions contrôlées par l'agent nécessitent l'utilisation de la même partie du corps de l'agent. Il est alors nécessaire de s'assurer que l'agent n'exécute qu'une des deux actions à la fois ou les fusionne. Si les valeurs

de contrôle des deux actions sont non nulles en même temps dans l'architecture, un mécanisme doit être capable de détecter le conflit dans la production de ces deux actions et de le résoudre en n'autorisant que la continuation d'une action à la fois.

8.2 Architectures existantes

Dans cette section, nous explorons les architectures d'agent existantes capables d'implémenter notre modèle. Nous avons mentionné dans la section 8.1.1 que seule une architecture incrémentale était capable d'implémenter notre modèle. En effet, une architecture d'agent traditionnelle requiert des modèles où perception, décision et action sont exécutées séquentiellement, l'agent n'interprétant pas l'action de l'utilisateur tant que ce dernier n'a pas fini son action. Nous nous intéressons ici uniquement aux architectures permettant une perception et une prise de décision en parallèle. De nombreux systèmes incrémentaux de dialogue ont été conçus depuis les années 1990. Beaucoup de ces architectures ont été créées par une seule équipe de recherche sans volonté de mutualiser les travaux réalisés par d'autres auteurs du domaine. Nous avons néanmoins comme objectif de montrer que notre modèle est intégrable dans une architecture d'agent avec des sous-modules d'interprétation et de génération du contenu, des sous-modules de gestion des émotions, ou de gestion des attitudes interpersonnelles. En ce sens, nous sommes intéressés par une architecture permettant d'intégrer des modèles réalisés par d'autres auteurs du domaine. ASAP (Kopp *et al.*, 2014), par le support du standard SAIBA (Kopp *et al.*, 2006), est ainsi l'architecture la plus adaptée pour l'implémentation de notre modèle. Néanmoins, ASAP, à l'instar de SAIBA, ne définit que les différentes étapes de traitement de l'architecture et les messages transmis entre les sous-modules de ces différentes étapes de traitement. Les contraintes présentées dans la section précédente, telle que la continuité de l'action ou la gestion des ressources, nécessitent la définition d'autres principes de conception de l'architecture. Pour la formulation de ces principes, nous avons choisi de nous inspirer de l'architecture Ymir. Ces architectures ayant été présentées dans le chapitre 2, nous ne les présentons pas de nouveau mais nous présentons dans les deux sous-sections suivantes en quoi ces deux architectures résolvent nos contraintes d'implémentation.

8.2.1 ASAP

Alors que nous traitons les grandeurs transitant dans l'architecture comme des grandeurs continues, ASAP traite des unités d'actions discrètes temporellement (phonèmes par exemple) plutôt que des valeurs continues. Néanmoins, contrairement aux architectures traitant de manière séquentielle la perception et la production d'action, ASAP rejoint nos problématiques d'implémentation en considérant que la perception et l'action s'exécutent en parallèle. L'architecture a de plus l'avantage

d'implémenter les langages FML proposé par Cafaro *et al.* (2014), BMLa proposé par (Kopp *et al.*, 2014) et PML tel que défini par Scherer *et al.* (2012). Ces langages ont été conçus comme des standards de communication entre sous-modules pour, respectivement, la génération d'actes communicatifs, le contrôle des actions motrices et la perception des actions de l'utilisateur. Le langage PML a l'avantage d'être compatible avec nos problématiques de perception continue du comportement. Ce langage ne fournit pas de contraintes temporelles entre deux éléments de perception, autorisant une perception temporellement continue, et offre la possibilité d'associer à un élément de perception un niveau de certitude. Le langage BMLa étend le langage BML pour introduire la capacité à moduler en continu les actions de l'agent, et à interrompre ou reprendre une action en cours, rendant ce langage compatible avec nos contraintes liées à la nature continue de l'action.

Intéressons-nous maintenant à la répartition de l'architecture en six modules de perception et d'action. L'architecture, telle que définie par Kopp *et al.* (2014), propose à la fois la possibilité d'implémenter des modèles de contrôle du comportement pro-actifs et réactifs. L'interprétation et le contrôle symbolique du comportement s'effectuent respectivement dans le module de *Function Interpretation* et l'*Intent Planner*. Le rôle du *Behavior Planner* pose plus de questions. Il assure en effet deux fonctions distinctes : il est chargé de sélectionner les actions « de surface » c'est-à-dire les actions concrètement exécutées dans l'environnement (que nous appelons commandes motrices selon les termes de Thórisson (2002) dans la suite) et il peut lui-même prendre des décisions sur le comportement de l'agent à partir des informations fournies par le *Behavior Interpreter*. Le contrôle du comportement de l'agent par le *Behavior Planner* est alors de nature réactive. Ainsi, selon Kopp *et al.* (2014), lors d'une interruption de l'utilisateur, l'agent détecte le fragment de parole de l'utilisateur dans le *Behavior Interpreter* et transmet l'information d'un fragment de parole détecté au *Behavior Planner* qui ordonne au *Behavior Realizer* d'arrêter de produire le comportement. Dans le cadre d'une implémentation de notre modèle, nous avons besoin à la fois de capacités pro-actives de l'agent et de capacités réactives. Notre modèle est en effet purement réactif : l'agent module ses signaux selon les informations perçues sur le comportement de l'utilisateur. Néanmoins, bien que nous ne modélisions pas la manière dont la motivation à parler de l'agent varie, nous considérons que cette variable est contrôlée du moins partiellement par des composantes pro-actives (génération d'un acte de langage par exemple). La capacité à générer à la fois ces deux types de comportement est donc un atout indispensable pour l'implémentation de notre modèle. Une clarification sur la manière dont sélection de l'action et gestion réactive du comportement sont réalisées est néanmoins nécessaire. Cette problématique rejoint la question du fonctionnement interne de chaque module, ce qu'ASAP ne précise pas. Nous avons choisi de reprendre certains principes proposés par Ymir pour définir le fonctionnement interne des modules.

8.2.2 Ymir

Notre modèle repose sur une perception et un contrôle de l'action s'effectuant en parallèle. À chaque signal produit par l'utilisateur est associée une fonction d'accumulation partielle indépendante des autres fonctions d'accumulation partielle, l'accumulation totale étant la somme de toutes ces accumulations partielles. De même pour le contrôle de l'action, chaque signal est indépendamment contrôlé par une équation différentielle différente. Pour toutes ces raisons l'emploi d'une architecture distribuée composée de modules spécialisés dans la perception ou le contrôle de chaque signal semble particulièrement adapté à nos problématiques. Une telle organisation faciliterait grandement l'ajout de nouveaux modules de perception ou d'action. Nous avons ainsi adopté certains principes d'Ymir, présenté dans le chapitre 2, page 38, pour l'implémentation des composantes de contrôle de l'action dans les différents modules de l'architecture ASAP. En plus d'être une architecture distribuée, Ymir propose la définition de *blackboards* pour la communication entre les différents modules de l'architecture. Ces *blackboards* rendent l'exécution de chaque module indépendant de l'exécution des autres modules et des événements ayant lieu dans l'environnement. L'exécution des modules est réalisée en parallèle et, peu importe la nature ou la disponibilité des données dans l'architecture, les modules s'exécutent à fréquence constante. Nous reprenons ces différents principes pour la conception de l'architecture.

L'architecture dispose d'un gestionnaire d'action (*action scheduler* en anglais) pour gérer le lancement de l'action. Ce module adresse la problématique de la gestion des ressources de l'agent et de la transcription des décisions d'action de l'architecture en commandes motrices réalisées par l'agent. Il implémente ainsi un mécanisme d'arbitrage entre les différentes décisions d'action générées en parallèle par les décideurs de l'architecture. Néanmoins, le gestionnaire d'action tel que proposé par Thórisson (1999) ne permet que l'exécution d'actions ayant un temps de début et de fin et ne gère pas la modulation en continu des paramètres de contrôle des commandes motrices ni l'interruption et la reprise des comportements générés dans l'architecture. Nous devons donc adapter le gestionnaire d'action actuel pour permettre la gestion de ces contraintes. Ymir est de plus une architecture traitant des données de nature événementielles et catégorielles plutôt que continues. Nous proposons ici de modifier le fonctionnement des modules pour inclure la possibilité de gérer des comportements continus.

8.3 Présentation de BeAware

Nous présentons maintenant notre architecture d'agent BeAware. Comme mentionné au début du chapitre, la conception d'une architecture n'étant pas l'objectif premier de notre travail, nous avons repris le plus possible les principes des ar-

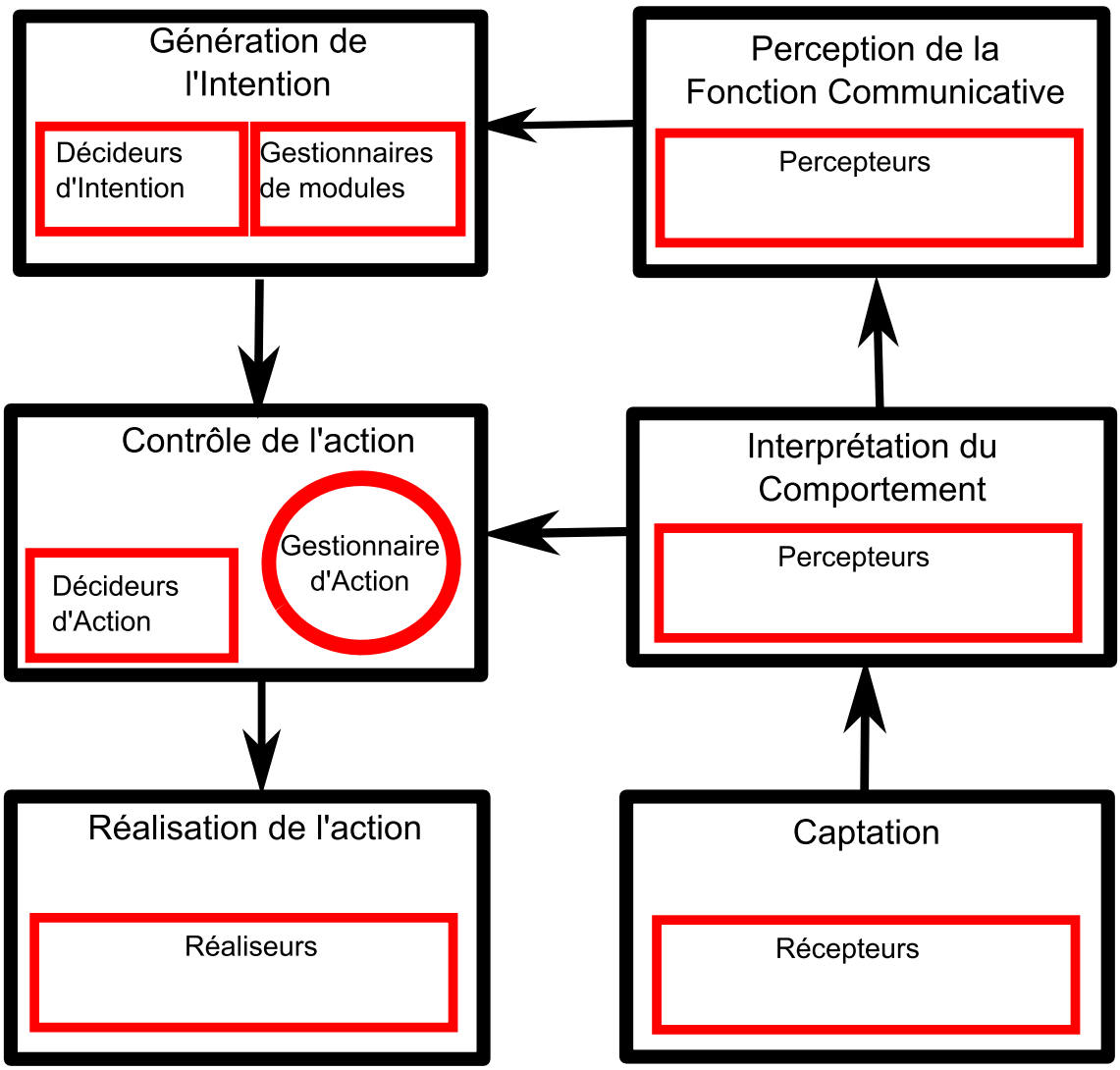


FIGURE 8.1 – Illustration de l'organisation globale de l'architecture. Les cadres rouges correspondent à l'implémentation de chaque sous-module dans l'architecture.

chitectures ASAP et Ymir pour la création de l'architecture. Néanmoins, ces deux architectures ne permettant pas en l'état l'implémentation d'un modèle de gestion continue et émergente du comportement, nous avons apporté des adaptations pour permettre l'implémentation, à la fois de modèles de contrôle continus et de modèles de contrôle événementiels.

8.3.1 Organisation de l'architecture

Nous reprenons la répartition en six modules provenant de l'architecture ASAP pour BeAware. Ces modules remplissent les mêmes fonctions que définies dans ASAP, et selon les principes d'Ymir, elles sont composées de sous-modules spécialisés s'exécutant indépendamment les uns des autres. La figure 8.1 décrit l'organisation des six modules de BeAware et décrit les sous-modules spécialisés que l'on retrouve dans chaque module. Le tableau 8.1 présente en détail les différents types de sous-modules présents dans l'architecture.

Nom	Module	Entrées	Sorties	Fonction
Récepteurs	Captation	Données des capteurs	Donnée de captation	Récupère les données d'entrées, normalise et publie une donnée traitable par les percepteurs
Percepteur	Perception du comportement Perception de la fonction communicative	Données de captation Données de perception	Donnée de perception	Prend en entrée une ou plusieurs données provenant des récepteurs ou des percepteurs de l'agent et produit une donnée de perception
Gestionnaire de module	Génération d'intention	Données de perception Données contextuelle	Activation-/désactivation de modules	Récupère en entrée les données de perception et active ou désactive les modules de l'architecture
Décideurs d'intention	Génération d'intention	Données de perception Donnée contextuelle	Acte communicatif	À partir de données perceptives et d'informations concernant le contexte du dialogue, génère un acte communicatif de l'agent
Décideur d'action	Génération d'action	Données de perception Acte communicatif	Commande d'action	À partir de données perceptives et d'actes communicatifs, génère une commande d'action
Gestionnaire d'action	Génération d'action	Commande d'action	Commandes motrices	Détermine à partir d'une commande d'action les actions motrices à réaliser et envoie les commandes d'action motrices aux réaliseurs
Réaliseurs	Réalisation d'action	Commande motrice		Se charge de lancer, interrompre, reprendre ou moduler selon la commande motrice reçue l'action motrice dont il est responsable

TABLE 8.1 – Présentation des différents modules de l'architecture.

Nous avons défini les différents types de sous-module selon les principes d'Ymir. Chaque sous-module de l'architecture peut être soit activé ou désactivé, excepté le gestionnaire d'action qui est tout le temps activé. Lorsqu'un sous-module est désactivé, sa fonction principale, calculant une valeur perceptive ou décidant d'une action à réaliser, n'est pas exécutée dans l'architecture.

En accord avec Ymir, nous distinguons les sous-modules de perception appelés les percepteurs, et les sous-modules de décision appelés les décideurs. Un sous-module récupérant les mouvements de la tête pour déterminer si l'utilisateur est en train de hocher la tête et un sous-module récupérant en entrée les hochements de têtes et le contenu verbal de la parole de l'utilisateur pour déterminer si l'utilisateur produit un *backchannel* sont des percepteurs de l'architecture. Nous distinguons par contre, contrairement à Ymir, les décideurs implémentés dans le module de génération d'intention (décideurs d'actes communicatifs) des décideurs du module de génération d'action (décideurs d'action). Les décideurs d'actes communicatifs génèrent des actes de communication de l'agent en fonction des données de perception fournies par les percepteurs du module de perception de la fonction communicative et de données contextuelles de l'interaction (statut social, ou objet pointé du doigt par un participant). Les paramètres des actes communicatifs reprennent les principes définis par Cafaro *et al.* (2014), chaque acte contient une information sur des contraintes temporelles liées à son exécution. Nous pouvons ainsi contraindre le bloc à être traité immédiatement, avant ou après un autre bloc. Les décideurs chargés de planifier l'énoncé de l'agent sont des exemples de décideurs d'acte communicatif. Les actes communicatifs générés par ces décideurs constituent des intentions communicatives de l'agent. Selon le principe de couplage de notre modèle (voir chapitre 7 par exemple), une motivation à ou une intention d'agir d'une certaine manière n'implique pas la réalisation effective et immédiate de l'action. De même selon les principes de continuité de l'action, une fois l'action lancée, elle est continuellement modulée par l'agent selon les signaux fournis par l'autre participant. Les décideurs d'action sont ainsi des intermédiaires entre la génération d'un acte communicatif de l'agent et le contrôle effectif des actions de l'agent en environnement virtuel. Ces sous-modules ont, d'une part, en entrée, un ou plusieurs actes communicatifs générés par les décideurs d'actes communicatifs et un ou plusieurs éléments de perception du comportement de l'utilisateur. En sortie, les décideurs d'action génèrent une commande d'action renseignant l'action à contrôler et la nouvelle valeur associée à cette action (par exemple hauteur de voix à 0.6, volume sonore à 1.0 comme présenté dans le chapitre 6). Pour clarifier la distinction entre les deux rôles mentionnés par Kopp *et al.* (2014), nous considérons qu'à chaque intention communicative de l'agent générée par un sous-module de planification d'intention, correspond un sous-module de contrôle de l'action chargé de contrôler et moduler l'action correspondant à cette intention communicative. La modulation du paramètre de contrôle

de l'action par le décideur conduit à la génération d'une requête d'action envoyée au gestionnaire d'action. Celui-ci récupère la requête d'action dans l'architecture et fait correspondre à cette requête des commandes motrices, à l'aide d'un lexique moteur, qu'il envoie ensuite aux réaliseurs concernés. Lors du choix des commandes motrices, le gestionnaire d'action vérifie que les ressources de l'agent associées à cette commande sont bien disponibles. Si les ressources ne sont pas disponible, le gestionnaire d'action cherchera d'autres commandes motrices remplissant le même rôle. Pour le fonctionnement global du gestionnaire d'action nous reprenons ainsi les principes d'Ymir. Nous verrons néanmoins plus en détail dans la section 8.3.3 le fonctionnement du gestionnaire d'action et montrerons en quoi certains principes de conception se distinguent d'ASAP et d'Ymir. Si le gestionnaire d'action n'est pas capable de trouver un ensemble minimal de commandes motrices lui permettant de produire une action, la production du comportement est annulée par le gestionnaire d'action.

Les gestionnaires de module reprennent les principes des décideurs non déclarés de l'architecture Ymir : selon l'état perçu du dialogue (agent engagé ou non dans la conversation, rôle locuteur ou auditeur de l'agent par exemple) ce sous-module active ou désactive les sous-modules liés à cet état du dialogue. L'existence de ces gestionnaires de module est une modification du mécanisme d'activation et de désactivation proposé originellement dans ASAP et correspond aux *covert deciders* de l'architecture Ymir. L'activation ou la désactivation de sous-module est en effet réalisée par l'envoi de messages entre modules dans ASAP : la décision d'activer ou désactiver des sous-modules est réalisée dans l'*Intent Planner*, puis envoyée au *Function Interpreter*, qui se charge d'envoyer à son tour le message au *Behavior Planner* et ainsi de suite. L'activation et la désactivation de sous-module nécessite donc plusieurs intermédiaires. Nous avons décidé de simplifier ce mécanisme en confiant la gestion de l'activation et de la désactivation des sous-modules à un seul sous-module. Ce sous-module possède une référence vers l'ensemble des sous-modules à activer ou désactiver. Il peut ainsi agir directement sur l'exécution de ces sous-modules. Lorsque l'agent passe du rôle de locuteur à auditeur, le gestionnaire de module désactive ainsi les sous-modules liés au rôle de locuteur (perception de prise de la parole du partenaire par exemple), et active les sous-modules liés au rôle d'auditeur. Nous détaillons maintenant, dans les sections suivantes, le fonctionnement des percepteurs, des décideurs, du gestionnaire d'action et des réaliseurs.

8.3.2 Percepteurs et décideurs

Le fonctionnement des percepteurs suit en grande partie les principes définis par Ymir. Chaque percepteur fonctionne en parallèle des autres percepteurs. Lors d'un cycle d'exécution, le percepteur va d'abord rechercher et lire les données de perception provenant d'un *blackboard* constituant ses entrées et, sur la base de ces entrées,

va produire en sortie une donnée de perception. Chaque donnée de perception a une ou plusieurs dimensions, les valeurs de chacune de ces dimensions pouvant être soit une chaîne de caractères représentant une donnée catégorielle ou un événement, soit une donnée numérique (variable réelle). Ce fonctionnement diffère ainsi des percepteurs d'Ymir par la nature de la donnée en entrée et en sortie de l'architecture, pouvant être non seulement catégorielle mais aussi numérique. En effet, à chaque exécution les percepteurs génèrent une nouvelle sortie, celle-ci étant stockée ensuite dans un *blackboard*. Nous nous distinguons ainsi de la nature événementielle de la prise de décision prônée par Thórisson (1999).

En accord avec les principes d'ASAP, une donnée provenant d'un percepteur ne peut pas être traitée par des modules de plus bas niveau que le module du percepteur générant la donnée. Ainsi, une donnée perceptive provenant d'un percepteur du module de perception de la fonction communicative ne peut être traitée par un sous-module du module de perception du comportement, tandis qu'un autre sous-module du module de perception de la fonction communicative pourra accéder à la donnée. Les données perceptives ont les mêmes attributs que défini par Scherer *et al.* (2012) pour le langage PML c'est-à-dire :

1. un nom,
2. une estampille temporelle,
3. le participant auquel se réfère la donnée de perception,
4. le module à partir duquel la donnée provient,
5. une valeur de certitude comprise entre 0 et 1,
6. d'autres paramètres optionnels spécifiques à la donnée de perception.

Pour l'accès des décideurs aux données de perception, nous respectons les principes énoncés par Kopp *et al.* (2014) : les décideurs du module de génération d'intention ne pourront avoir accès qu'aux données perceptives du module d'interprétation de la fonction communicative et les décideurs du module de contrôle de l'action ne pourront avoir accès qu'aux données perceptives du module de perception du comportement. Les décideurs utilisent ces données perceptives pour la génération d'un acte communicatif pour les décideurs d'actes communicatifs ou pour contrôler l'action à réaliser pour les décideurs d'action. Plus précisément, durant un cycle d'exécution, chaque décideur récupère d'abord dans les *blackboards* correspondants les données lui servant à contrôler l'acte communicatif ou l'action à réaliser puis calcule sur la base de ces données la valeur des paramètres de l'acte communicatif ou de la commande d'action.

En accord avec le langage FML tel que formulé par Cafaro *et al.* (2014), chaque acte communicatif généré par un décideur du module de génération d'intention dispose d'un ensemble de données déclaratives composées d'informations sur chaque participant engagé dans une interaction avec l'agent. Ces données déclaratives comprennent pour l'instant le nom des participants et les différentes conversations (*floor*)

dans lesquelles est engagé l'agent. Certaines informations déclaratives mentionnées par Cafaro *et al.* (2014), comme le niveau de relation entre l'agent et le participant, la personnalité du participant, son âge ou son genre n'ont pas encore été implémentées. Un acte communicatif a également un ou plusieurs blocs spécifiant un comportement à exécuter par l'agent (prendre le tour, laisser le tour par exemple). Chaque bloc a un identifiant, la référence du participant auquel est adressé le comportement, la catégorie (acte de langage, gestion du tour de parole...) à laquelle appartient l'acte et le type d'acte (affirmer, ordonner pour les actes de langage ou prendre le tour, laisser le tour pour la gestion du tour de parole par exemple). Nous étendons le langage FML existant pour inclure comme paramètre additionnel de chaque bloc un niveau de motivation à réaliser le comportement spécifié par le bloc. Afin de garder une compatibilité avec le standard SAIBA ce paramètre reste optionnel : si la valeur de motivation n'est pas spécifiée, les sous-modules considèrent que la motivation de l'agent est maximale. Le concept de motivation a une définition similaire à ce que nous avons proposé dans le chapitre 6. Nous reprenons les différents paramètres de contrôle de l'action proposés par BML et son extension BMLa pour définir une commande d'action. La requête d'action a alors les attributs suivants :

1. Un identifiant associé au personnage virtuel destinataire de ce bloc.
2. Un attribut « Composition », renseignant les contraintes temporelles de la commande par rapport aux actions précédentes lancées dans le réaliseur. La valeur de ce paramètre peut être soit MERGE , indiquant que la commande sera réalisée en parallèle des actions actuellement exécutées, soit APPEND, indiquant que la commande d'action sera traitée lorsque les autres actions se seront arrêtés soit REPLACE, ayant pour effet d'annuler l'exécution de toutes les actions en cours avant d'exécuter la commande.
3. Un attribut « Commande », provenant du langage BMLa et précisant si le réaliseur doit lancer (ACTIVATE), moduler (VALUECHANGE), interrompre (INTERRUPT), reprendre (RESUME) ou arrêter l'action (STOP).
4. Une durée de démarrage et de fin de l'action dans le cas d'actions ponctuelles.
5. Des contraintes de synchronisation reprenant la syntaxe définie dans le langage BMLa.
6. Une référence vers les unités d'action après lesquelles (CHUNKAFTER) et avant lesquelles (CHUNKBEFORE) l'action est exécutée.

Afin de garantir la transmission des données entre les sous-modules de l'architecture, des *blackboards* sont associés à chaque module de l'architecture, de sorte que seuls les sous-module du module auquel est associé le *blackboard* peuvent lire les données de ce *blackboard*. Les percepteurs ne peuvent de même publier leurs données que dans les *blackboards* du module auquel ils appartiennent, dans les *blackboards* du module au-dessus de celui auquel ils appartiennent et dans le *blackboard* des décideurs situés au même niveau qu'eux : les percepteurs du module d'interprétation

d'action pourront ainsi publier soit dans le *blackboard* du module de gestion des actions, soit dans le *blackboard* du module de perception des actions, soit dans le *blackboard* du module d'interprétation des fonctions communicatives. Les percepteurs du module d'inteprétation des fonctions pourront, quant à eux, publier dans le *blackboard* du module d'interprétation des fonctions et dans le *blackboard* du module de génération d'intention. Pour les décideurs, les décideurs d'actes communicatifs publieront uniquement dans le *blackboard* du module de contrôle de l'action et les décideurs du module de contrôle de l'action ne publieront dans aucun *blackboard*.

Ces *blackboards* sont implémentés selon les mêmes principes que dans l'architecture Ymir. Lorsqu'un sous-module de perception ou de décision d'intention génère une sortie, il publie la donnée dans un *blackboard*, les sous-modules ayant besoin de cette donnée lors de leur exécution effectueront alors une requête auprès du *blackboard* stockant cette donnée. Ce système nous permet de rendre l'exécution de chaque sous-module indépendant de l'exécution des autres sous-modules, à l'instar d'Ymir. Toute donnée possédant un nom, une estampille temporelle et un ensemble optionnel de valeurs peut être stockée dans un *blackboard*. La forme de la donnée diffère de ce qui est proposée dans l'architecture Ymir où celle-ci dispose d'un nom, d'un état, et d'une estampille temporelle. Pour illustrer ce principe de communication par *blackboard* la figure 8.2 résume la manière dont les sous-modules communiquent entre eux et les données échangées entre les modules.

8.3.3 Gestion des commandes motrices

Les variables d'actions modulées en sortie par les décideurs sont transmises au gestionnaire d'action sous forme de commandes d'action. À la réception d'une commande d'action, le gestionnaire d'action commence par vérifier si une action du même nom est déjà exécutée par l'agent. S'il voit qu'une action est démarrée et que la commande reçue lui demande de déclencher une action (commande ACTIVATE) il ne prend pas en compte cette commande. Le gestionnaire d'action va ensuite faire la correspondance entre la commande d'action et des actions motrices multimodales de l'agent. Cette correspondance est effectuée par l'intermédiaire d'un « lexique moteur » que nous détaillons dans le paragraphe suivant.

Dans Ymir, le lexique moteur est représenté sous forme d'arbre. Les sommets et nœuds non-terminaux de cet arbre, appelés des schèmes d'action, représentent soit des actions multimodales, soit des actions pouvant être réalisées par l'agent de plusieurs manières. Chaque schème d'action dispose de successeurs pouvant soit être un schème d'action, soit être un schème moteur. Ces schèmes moteurs constituent, eux, les feuilles du lexique moteur et correspondent aux commandes motrices exécutées par l'agent. Dans BeAware, chaque schème moteur fournit des renseignements sur les parties du corps de l'agent utilisées par la commande motrice correspondante, sur l'exécution actuelle ou non de la commande motrice par un réaliseur et une liste de

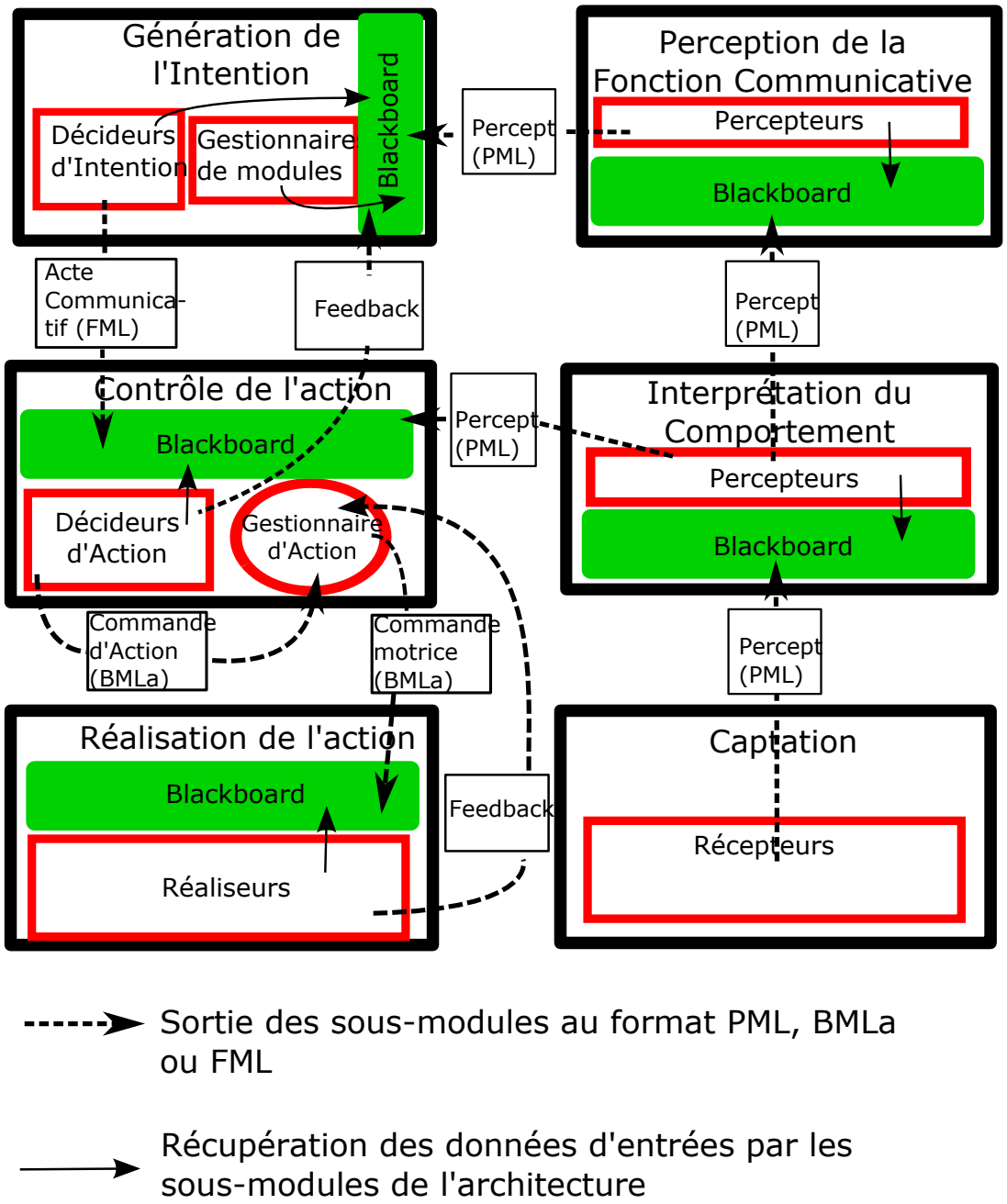


FIGURE 8.2 – Schéma illustrant la communication entre les sous-modules de l'architecture.

paramètres associée à la commande motrice (vitesse de gestuelle, hauteur de voix, volume sonore...).

Lorsque le gestionnaire d'action parcourt cet arbre à la recherche d'un ensemble d'actions motrices associé à une commande d'action, il commence d'abord par chercher dans l'arbre l'action multimodale du même nom que le nom mentionné dans la commande d'action. Il parcourt le lexique moteur, pour trouver un ensemble de schèmes moteurs permettant d'exécuter la commande d'action reçue en entrée. La sélection des schèmes moteurs est un processus dépendant de la nature de la commande d'action, de ses paramètres, de la disponibilité des ressources corporelles de l'agent ou encore de l'exécution des actions motrices ou non par d'autres actions. La sélection peut donc être spécifique à chaque schème d'action de l'arbre. Dans BeAware, chaque schème d'action dispose d'un sélecteur d'action propre définissant la politique de sélection des commandes motrices de l'agent pour ce schème d'action particulier. D'autres sélecteurs sont génériques à plusieurs actions différentes. Nous avons ainsi défini le sélecteur ET, sélectionnant un ensemble de commandes motrices que si chaque élément de cet ensemble est disponible et le sélecteur OU, choisissant entre plusieurs commandes motrices alternatives, de sorte que si une commande motrice n'est pas disponible, une autre commande motrice disponible pourra être choisie. La production de *backchannels* constitue un exemple de choix entre plusieurs alternatives, un auditeur pouvant choisir entre produire un hochement de tête ou une vocalisation. Il est aussi possible d'implémenter des chaînes de sélections d'action. Par exemple, un sélecteur ET peut être associé à deux sélecteurs OU : l'action requiert alors deux actions motrices qui peuvent être chacune choisies entre plusieurs alternatives. La figure 8.3 montre un exemple de lexique moteur avec un schème d'action« montrer son désaccord », disposant de trois sélecteurs ET et OU chainés.

Une fois les actions motrices choisies par le gestionnaire d'action, ce dernier publie les commandes motrices dans un *blackboard* de commandes motrices. À l'exécution, les réaliseurs commenceront par chercher dans ce *blackboard* si une nouvelle commande motrice correspondant à l'action contrôlée par le réaliseur a été publiée par le gestionnaire d'action. Nous reprenons la même syntaxe et les mêmes attributs que la commande d'action pour la commande motrice.

8.3.4 Réaliseurs d'action

Dans BeAware, les réaliseurs d'action ont, comme les autres sous-modules de l'architecture, une boucle d'exécution indépendante des autres sous-modules de l'architecture. À chaque itération de la boucle d'exécution, le réaliseur récupère la dernière commande motrice concernant l'action qu'il supervise. Si cette commande est différente de la dernière commande exécutée, le réaliseur se charge d'appliquer les changements spécifiés par cette commande. Il vérifie que la commande spécifiée

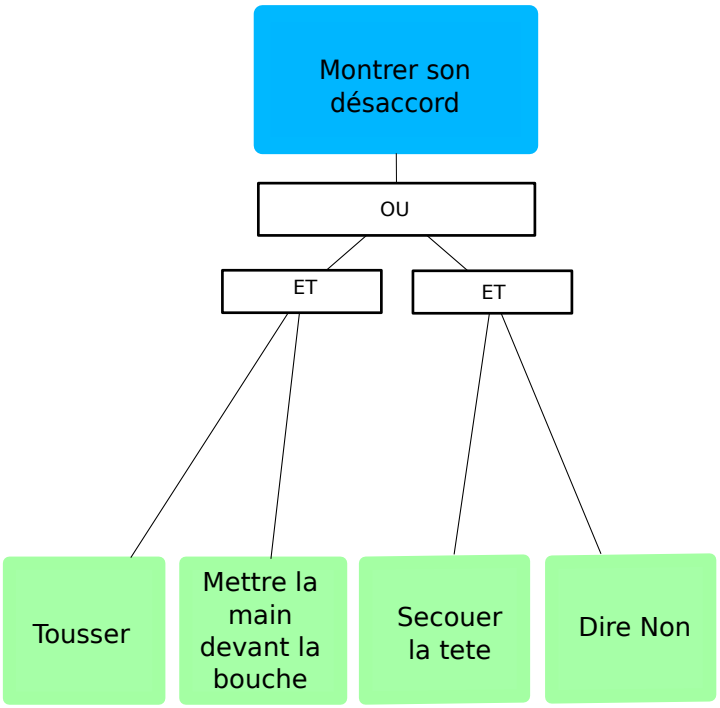


FIGURE 8.3 – Exemple d'arbres avec deux schèmes d'actions et six schèmes moteurs. Les blocs avec fond rempli représentent les schèmes d'action et les schèmes moteurs de l'arbre, les blocs avec un fond blanc représentent les sélecteurs de l'architecture.

dans la commande motrice (*PREPLAN*, *ACTIVATE*, *VALUECHANGE*, *STOP*) est cohérente avec l'état courant de l'action. Ainsi, la commande *PREPLAN* ne sera exécutée par le réaliseur que si l'action n'est pas exécutée, la commande *ACTIVATE* ne sera exécutée que si l'action n'est pas exécutée ou a été pré-planifiée, et la commande *STOP* que si l'action est actuellement exécutée dans l'environnement, reprenant les principes des langages BML et BMLa. Une fois les changements appliqués, le réaliseur met à jour l'état de l'action. Cette valeur d'état de l'action peut prendre les états suivants en accord avec le standard SAIBA tel que présenté par Kopp *et al.* (2006) :

1. NOTRUNNING indique que l'action n'est pas encore démarrée,
2. PREPLAN indique que l'action a été pré-planifiée,
3. PENDING, indique que le réaliseur a exécuté une commande motrice mais est toujours en attente de la prise en compte de la modification de l'action par les actionneurs de l'agent,
4. START indique que l'action est en cours d'exécution,
5. READY, marque la fin de la phase de préparation d'une action,
6. STROKE_START, marque le début de la phase où l'action de l'agent porte le plus de signification,
7. STROKE, marque la phase principale de réalisation de l'action,

8. STROKE_END, marque la fin de la phase où l'action de l'agent porte le plus de signification,
9. RELAX, marque la transition, de l'action vers l'état de repos,
10. INTERRUPTED indique que l'action a été interrompue,
11. END indique que l'action a été terminée,

Nous avons repris l'ensemble des états proposés par SAIBA et ASAP. Nous avons néanmoins défini l'état « PENDING », renseigné par le gestionnaire d'action ou le réaliseur lorsque ces sous-modules envoient une commande d'action (respectivement à un réaliseur ou à un actionneur) et n'ont pas encore reçu de retours sur l'exécution de la commande.

Les différents états présentés ci-dessus peuvent ou non s'appliquer à une commande motrice, tel que défini par SAIBA. Ainsi, si la commande motrice n'a pas de phase de préparation, l'état *READY* sera le même que l'état *START* et si la commande n'a pas de phase de relâche, l'état *STROKE_END* coïncidera avec l'état *RELAX*. Si l'action est ponctuelle et ne peut être interrompue, l'action ne pourra pas être dans l'état *INTERRUPTED*. Le réaliseur surveille la progression de l'action dans l'environnement virtuel et met à jour l'état courant de l'action. Après l'exécution d'une commande, le réaliseur met à jour l'état de l'action à *PENDING*. Lorsque le changement est effectif, il met à jour l'état de l'action et envoie au gestionnaire d'action un *feedback* concernant la modification de l'état de l'action. Une donnée de *feedback* comprend un nom faisant référence à la commande motrice exécutée, une estampille temporelle et l'état courant de l'action.

À partir du *feedback* envoyé par le réaliseur, le gestionnaire d'action met alors à jour l'état de la commande motrice et met à jour en conséquence la commande d'action d'où provient la commande motrice. Afin d'avoir un retour sur la progression des actions, les décideurs ont accès à la liste des actions en cours dans l'environnement. Ils peuvent ainsi envoyer à leur tour une donnée de *feedback* aux décideurs d'actes communicatifs du module de contrôle des intentions de l'agent, fournissant un état d'avancement des actes communicatifs provenant des décideurs du module de génération des intentions.

8.4 Conclusion

Nous avons présenté au cours de ce chapitre BeAware, une architecture d'agent reprenant en grande partie les architectures Ymir (Thórisson, 1999) et ASAP (Kopp *et al.*, 2014) et supportant l'implémentation de notre modèle de gestion du tour de parole pour une interaction temps-réel entre un agent et un utilisateur. Nous avons tout d'abord présenté les qualités que doit avoir une architecture d'agent implémentant notre modèle. Sur la base de ces qualités et de notre volonté de rendre notre modèle intégrable avec une grande variété de sous-modules de contrôle du

comportement de l'agent, nous avons identifié l'architecture ASAP comme candidate potentielle pour l'implémentation de notre modèle. Néanmoins ASAP a des manques concernant la manière dont les différents sous-modules de perception, de contrôle de l'action et de réalisation de l'action doivent être implémentés. Nous avons résolu certaines problématiques d'implémentation des sous-modules de contrôle du comportement grâce à Ymir. Ymir est une architecture où les données traitées sont de nature événementielle et non continue. Nous avons donc proposé notre propre architecture inspirée d'ASAP et d'Ymir permettant l'implémentation de sous-modules de contrôle continus du comportement. Par son implémentation quasi-complète du standard SAIBA cette architecture permet l'interconnexion avec n'importe quel module comportemental, incrémental ou non, implémentant le standard SAIBA. Nous rendons donc notre module de gestion du tour de parole réutilisable avec d'autres modules (gestion du contenu, gestion des émotions entre autres ...). Nous détaillons dans le chapitre 10 la manière dont nous avons implémenté notre modèle dans cet architecture et nous démontrons son fonctionnement dans un scénario d'interaction entre l'utilisateur et l'agent.

Troisième partie

Validation du modèle

Chapitre 9

Calibration du modèle

Nous avons présenté dans les chapitres 6 et 7 des exemples de situations d'interaction entre deux agents vérifiant les sept hypothèses présentées en section 6.1 du chapitre 6. Nous rappelons ces hypothèses dans le tableau 9.1.

Les simulations présentées dans ces chapitres sont théoriques, ne reproduisant en rien les variations de signaux, les durées de transition et de conflit observées dans les interactions humaines. Nous souhaitons maintenant aller vers une interaction utilisateur-agent où les deux participants s'échangent le tour en faisant varier leur volume sonore et leur hauteur de voix. Pour aller vers une interaction naturelle où l'utilisateur et l'agent coordonnent leurs échanges de tour, nous devons nous assurer que l'agent soit capable de reproduire les comportements observés au cours d'interactions humaines. Pour une interaction réussie avec un utilisateur, l'agent et l'utilisateur doivent moduler leurs signaux de sorte de respecter les durées de transition et de conflit que l'on peut observer dans des interactions humaines. Cela implique que l'agent ait la capacité à percevoir et à réagir aux variations de signaux de la même manière qu'un participant humain le ferait, et la capacité à faire varier ses signaux de sorte que l'utilisateur interprète correctement ses motivations.

Nous avons aussi choisi d'implémenter les équations de contrôle de l'agent à partir de données récoltées d'interactions dialogiques humaines. Nous avons ainsi élaboré un protocole expérimental destiné à mesurer la variation des durées de transition et de conflit ainsi que les variations des signaux prosodiques des participants. En complément, par ce protocole, nous souhaitions valider certaines hypothèses de conception du modèle présentées dans le tableau 9.1. Ce protocole expérimental est présenté en section 9.1. À partir des données provenant de l'expérimentation, nous avons déduit une implémentation du modèle que nous présentons en section 9.2. Nous avons ensuite vérifié la capacité de deux agents à se coordonner afin de reproduire les durées de transition et de conflit observées dans le cadre de notre expérimentation. Nous présentons enfin les simulations agent-agent résultantes en section 9.3.

Hypothèse 1	L'occurrence d'un changement de tour peut autant être à l'initiative de l'auditeur que du locuteur.
Hypothèse 2	L'occurrence et la durée d'une transition ou d'un conflit est dépendante des motivations à parler des deux participants.
Hypothèse 3	Un participant perçoit les motivations de parler de l'autre en interprétant l'ensemble des signaux produits par l'autre.
Hypothèse 4	Un participant coordonne sa production d'action afin de signaler sa motivation de parler ou non.
Hypothèse 5	La perception du comportement de prise ou d'abandon du tour se fait de manière continue par accumulation d'indices provenant de toutes les actions produites par l'autre.
Hypothèse 6	Les comportements des participants au cours de l'interaction sont émergents de leur interaction mutuelle.
Hypothèse 7	Les actions de l'agent sont directement modulées par, d'une part, la perception du comportement de son partenaire et d'autre part par sa propre motivation à parler.

TABLE 9.1 – Rappel des hypothèses de conception du modèle présentées chapitre 6.

9.1 Analyse du tour de parole dans les interactions humaines

9.1.1 Questions de recherche et hypothèses conceptuelles

Le protocole expérimental a été conçu pour répondre à une question de recherche et des hypothèses conceptuelles que nous avons formulées à partir des hypothèses de conception du modèle. Nous présentons dans cette première sous-section ces questions de recherche et les hypothèses conceptuelles testées au cours du protocole expérimental.

Par les hypothèses 3 et 4 présentées dans le tableau 9.1, nous avons affirmé que les participants interprétaient et variaient leurs signaux sur tous les canaux de communication verbaux et non-verbaux à leur disposition pour se coordonner. Si c'est le cas, les participants interprètent et coordonnent conjointement les informations verbales, prosodiques et non-vocales pour s'échanger la parole dans toutes les situations et pour tous les participants. Comme corolaire de ce principe nous avons observé dans une simulation agent-agent (chapitre 7) que lorsque l'on diminuait le nombre de signaux interprétés par l'agent on observait une adaptation des agents qui modifiaient leurs variations de signaux. Ainsi, lorsque l'information du regard était manquante, les agents avaient tendance à accentuer leurs variations de volume sonore et de hauteur de voix pour se coordonner. Cette capacité montre l'intérêt d'un modèle s'appuyant sur le couplage entre les deux agents pour gérer les échanges de paroles. Si un tel couplage existe entre les participants humains, nous devrions

observer une adaptation similaire des participants en enlevant les informations liées aux signaux non-vocaux. Nous formulons alors la question de recherche suivante :

Question 1. *Les variations de signaux prosodiques observées lors de transitions de tour et lors de conflits et les durées de transitions et de conflit sont-elles différentes dans les situations où les participants se voient par rapport aux situations où les participants ne se voient pas ?*

En lien avec cette question de recherche, nous émettons les deux hypothèses conceptuelles suivantes :

HC 1. *On observe toujours une coordination lorsque les participants ne se voient pas.*

HC 2. *La coordination est dégradée lorsque les participants ne se voient pas.*

Pour valider l'hypothèse conceptuelle 1 nous cherchons à confirmer l'hypothèse suivante :

HO 1. *Plus de 50 % des transitions se font avec un moment de silence de moins d'une seconde entre les tours des deux participants lorsque les participants ne se voient pas.*

Pour valider l'hypothèse conceptuelle 2 nous souhaitons confirmer l'hypothèse suivante :

HO 2. *On observe une variation des durées de transition entre la condition où les participants se voient et la condition où les participants ne se voient pas. Le nombre de transitions comportant des silences longs et de transitions avec recouvrement est plus grand dans le cas où les participants ne se voient pas par rapport à la condition où les participants se voient.*

Protocole expérimental

Durant l'expérimentation les participants dialoguent pendant six minutes : ils se voient pendant trois minutes, et sont cachés l'un de l'autre pendant trois autres minutes. En divisant l'interaction en deux conditions, nous avons cherché à mesurer l'effet de l'absence de signaux non-vocaux sur les échanges de tour entre les participants avec pour objectif de tester les hypothèses conceptuelles 1 et 2. Pour éviter un effet de l'ordre des conditions sur les résultats expérimentaux nous avons contrebalancé les conditions. Afin d'éviter des biais liés au sujet de la conversation, toutes les dyades étaient chargées d'échanger sur le même sujet. Nous avons choisi de faire dialoguer les participants selon un scénario de survie, de tels scénarios approximant, selon Burgoon *et al.* (2000), les caractéristiques d'une conversation normale. Il était ainsi présenté à chaque paire le scénario de survie suivant : « Vous êtes sur un bateau en train de couler, vous et votre partenaire vous préparez à embarquer sur un

bateau de sauvetage comportant déjà un stock d'eau vous permettant de subsister plusieurs semaines à deux ».

Il est de plus donné à chaque participant une indication différente sur sa situation actuelle. Ainsi pendant l'interaction, un des participants pense que le bateau est proche de la côte tandis l'autre participant pense que le bateau est loin de la côte. Il est alors suggéré, dans la consigne, au participant pensant qu'il est proche de la côte, de prendre des objets permettant de se faire repérer ou d'atteindre la côte le plus rapidement possible. Il est ainsi proposé dans la consigne de prendre les trois objets suivants :

— un GPS,
— une fusée de détresse,
— des rames.

Pour le participant persuadé d'être loin de la côte, il lui est proposé de privilégier des objets favorisant la survie des participants, c'est-à-dire :

— vingt rations de nourriture,
— des accessoires de pêche,
— des couvertures.

Au cours des six minutes de l'interaction, les participants ont pour objectif de se mettre d'accord sur trois objets parmi les six proposés aux deux participants. Ce scénario a été conçu afin de permettre l'observation à la fois de moments de consensus entre les participants et de moments de conflits, générant une grande variété de situations communicatives liées au tour de parole.

13 dyades ont en tout participé à cette expérimentation. Les participants ont tous répondu individuellement à l'appel à participant. Nous nous sommes alors chargés de former les dyades de sorte de remplir les critères mentionnés ci-dessous.

1. Avoir pour langue maternelle le français.
2. Éviter des différences dans le statut social des participants pouvant amener à l'établissement d'attitude de soumission ou de dominance entre les participants. Nous avons fait en sorte que dans la majorité des cas des étudiants dont le niveau d'étude était inférieur au doctorat interagissaient avec des étudiants ayant un niveau d'étude similaire, des doctorants évitent d'interagir avec des enseignants-chercheurs.
3. Éviter des dyades mixtes homme-femme.
4. Faire en sorte que les participants ne se connaissent pas.

Malgré tout nous n'avons pas eu d'autres choix que de former quatre dyades mixtes, et quatre dyades dont le statut des participants différaient nous laissant avec 5 dyades remplissant les critères ci-dessus. L'âge des participants était compris entre 18 et 35 ans.

9.1.2 Résultats expérimentaux

Nous présentons dans la suite, les mesures de durées de transition et de variations de signaux que nous avons récoltées à partir des interactions humaines. Nous présentons dans une première sous-section la manière dont nous avons mesuré le volume sonore et la hauteur de voix de chaque participant. Nous présentons ensuite notre analyse des durées de transition que nous avons mesurées pour la condition où les participants se voient et pour la condition où les participants ne se voient pas. Nous présentons enfin la mesure des variations des signaux prosodiques dans les deux conditions.

Extraction de la prosodie

Lors des interactions dialogiques, chaque participant disposait d'un micro-casque dans lequel il parlait. Le micro-casque était disposé de sorte de ne pas affecter la perception de la voix de l'autre participant tout en gardant une distance constante entre le participant et le micro, écartant la possibilité d'une variation du volume sonore due à la variation de la distance entre le participant et le micro. L'enregistrement de la voix de chaque participant était effectué séparément. Pour aligner temporellement les enregistrements des deux participants, l'expérimentateur produisait, pour signaler le début de l'interaction, un son perçu par les micros des deux participants. Les enregistrements étaient ensuite alignés manuellement de sorte que le son produit par l'expérimentateur dans les deux enregistrements soit placé à $t = 0$.

À partir de ces enregistrements, pour chaque participant, le volume sonore et la hauteur de voix ont été extraits en utilisant l'outil d'analyse linguistique Praat (Boersma, 2002) à un taux d'échantillonnage de 10 ms. Le volume sonore a été extrait en utilisant la méthode d'interpolation cubique proposée par l'outil et la hauteur de voix a été extraite en utilisant la méthode d'auto-corrélation. Pour chaque algorithme, les paramètres par défaut de Praat ont été utilisés. La figure 9.1 illustre un exemple de variation de volume sonore extrait de l'enregistrement de l'interaction, ainsi qu'une transcription de l'énoncé prononcé par le participant.

L'algorithme de détection de la hauteur de voix proposé par Praat extrait la valeur de fréquence fondamentale des voyelles et consonnes voisées (produites par une vibration de la corde vocale) prononcées par les participants, les segments où le participant prononce une consonne non voisée (tel que p ou t) et les segments où il ne parle pas sont, eux, considérés comme des segments non voisés. La figure 9.2 illustre un exemple de variation de hauteur de voix avec une transcription de l'énoncé prononcé par le participant en bas de la figure. On constate sur cette figure que lors de la prononciation des consonnes non voisées (p ou s) aucune valeur de hauteur de voix n'est attribuée par l'algorithme.

Néanmoins, les algorithmes actuels de détection de hauteur de voix génèrent des faux positifs, considérant parfois les segments où l'utilisateur prononce une consonne

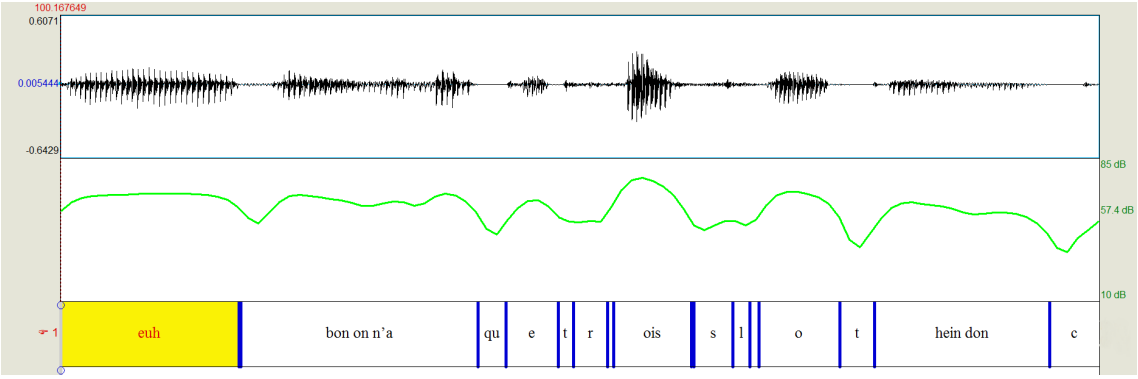


FIGURE 9.1 – Exemple d'une variation de volume sonore extraite d'un enregistrement. En haut de la figure est représentée la forme d'onde, la courbe verte en dessous représente le volume sonore et la transcription de l'enregistrement est représentée en bas de la figure.

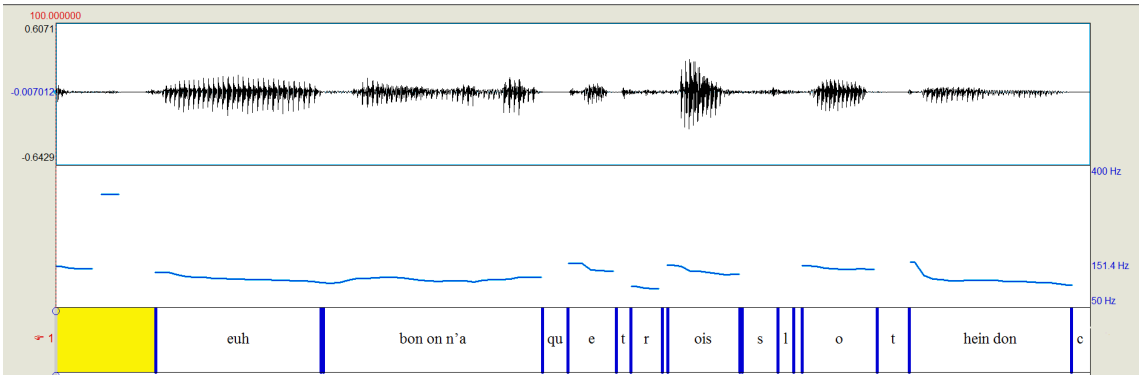


FIGURE 9.2 – Exemple d'une variation de hauteur de voix extraite d'un enregistrement.

non voisée comme un segment parlé. La valeur de hauteur de voix détectée pour ces segments est alors anormalement élevée ou basse. Aussi pour supprimer au maximum la détection erronée de ces segments, nous avons fixé un intervalle des valeurs que pouvait prendre la valeur de hauteur de voix des participants de sorte que lorsque la hauteur de voix n'est pas comprise dans cet intervalle le segment est considéré comme non voisé. Pour l'ensemble des participants, nous avons fixé la valeur minimale de hauteur de voix à 50 Hz et la valeur maximale à 400 Hz. Cela permet de supprimer un grand nombre de faux positifs dans les valeurs de hauteur de voix, tout en s'assurant qu'aucun segment voisé n'est supprimé de l'analyse. Certaines valeurs erronées n'ont néanmoins pas été supprimées par l'application de cet intervalle. Une raison à cela réside dans l'intervalle de détection utilisé, volontairement étendu pour prendre en compte la variabilité dans les valeurs de hauteur de voix de chaque participant. Pour optimiser ces valeurs nous avons repris la méthodologie utilisée par Oertel *et al.* (2011). Nous calculons à partir de la courbe extraite les déciles des valeurs de hauteur de voix du participant. Nous déterminons alors de nouveau les limites des valeurs de hauteur de voix de sorte que la valeur minimale détectable de hauteur de voix soit égale à 0.83 fois la valeur du quinzième décile et la valeur maximale soit égale à 1.92 fois la valeur du soixante-cinquième décile. Nous optimisons ainsi la suppression de faux positifs en adaptant l'intervalle de détection de hauteur de voix aux valeurs de chaque participant.

Les valeurs de volume sonore sont, quant à elles, très sensibles aux bruits ayant lieu dans l'environnement du participant. Pour éliminer l'influence de ces bruits, nous nous basons sur la distinction entre les segments voisés et les segments non voisés de l'algorithme de détection de hauteur de voix. Les valeurs de volume sonore sont ainsi forcées à 0 lorsque la hauteur de voix est elle-même à 0, indiquant un segment non voisé, et sont gardées à leur valeur actuelle lorsque la hauteur de voix est supérieure à 0. La figure 9.3b montre un exemple de profil de volume sonore filtré selon cette méthode. Les résultats montrent une figure entrecoupée de moments de silence et de moments où une voix du participant est détectée. Cela complique l'extraction des valeurs de variation de volume sonore et de variation des valeurs de hauteur de voix ainsi que le calcul des durées de transition. En effet, les moments sans voix correspondent à deux situations différentes, un moment où le participant effectue une pause dans son discours et un moment où le participant prononce une consonne non voisée. Pour la détection de la fin de tour, et la détection de conflit, nous devons savoir exactement quand les participants finissent leurs tours ou effectuent une pause. Nous souhaitons avoir un segment continu de parole lorsque le participant produit un énoncé sans effectuer de pause. De ce que nous avons pu observer des enregistrements des interactions, la durée de prononciation d'une ou plusieurs consonnes non voisées est au maximum de 400 ms. En reprenant la méthode de Hjalmarsson (2011) et Gravano et Hirschberg (2011), nous supprimons les

pauses de 400 ms en effectuant une interpolation linéaire entre les valeurs précédant la micro-pause et les valeurs suivant la micro-pause. Nous obtenons alors une suite de segments continus de parole entrecoupés de pauses et de fins de tour de la part des participants similaires aux unités inter-pausales traitées par Gravano et Hirschberg (2011) et Hjalmarsson (2011). Nous illustrons sur la figure 9.3 un exemple de traitement d'un profil prosodique en partant du volume sonore provenant de Praat puis filtré en tenant compte des segments voisés et non voisés et interpolé. Une interpolation similaire a été effectuée pour la hauteur de voix.

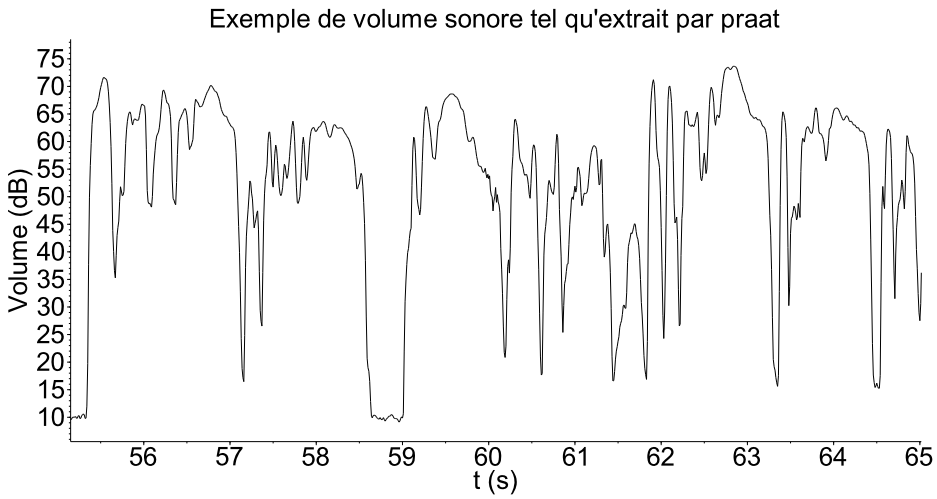
Une fois les données de volume sonore et de hauteur de voix extraites et traitées, nous avons segmenté les enregistrements en tours de parole. Pour cette segmentation, nous avons conçu un algorithme de détection automatique des tours de parole. Cet algorithme utilise uniquement les valeurs de volume sonore des deux participants et distingue à chaque moment de l'interaction trois cas de figure : deux cas correspondant à une situation où l'un des participants a le tour et un cas correspondant à une situation où personne ne parle. La détection des tours de parole se fait comme suit :

— si le participant est le locuteur courant, que son niveau sonore est non nul et que le niveau sonore de son partenaire est nul il reste locuteur courant,
— si son partenaire se met à parler pendant plus d'une seconde un changement de tour a lieu, le locuteur précédent devient l'auditeur courant et l'auditeur précédent devient le locuteur courant,
— si les deux participants parlent en même temps pendant plus de 300 ms, un conflit de parole est détecté,
— si personne ne parle pendant plus d'une seconde, un silence « gênant » est détecté.

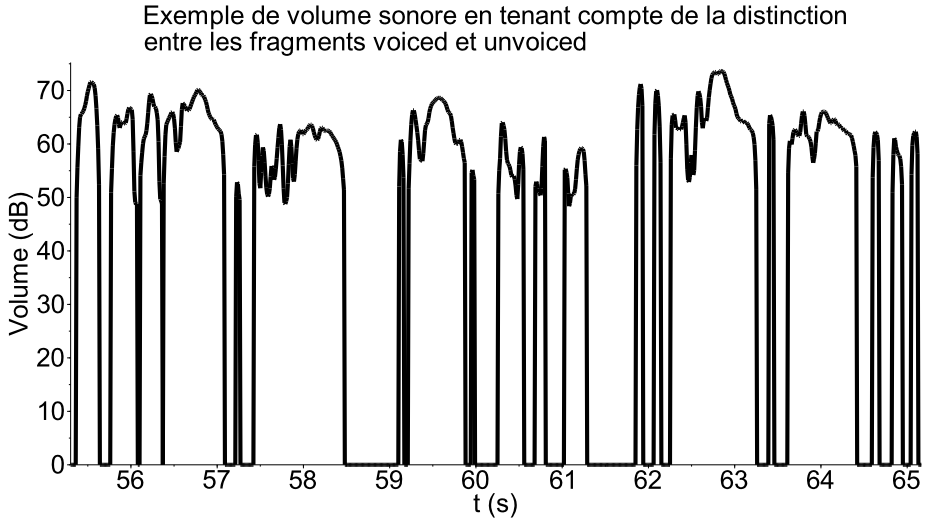
Le seuil d'une seconde pour détecter un changement de tour sert, lui, à éviter de considérer un *backchannel* comme une prise de tour de la part du locuteur courant. Nous fixons une durée de 300 ms pour la détection du conflit afin d'écarter des recouvrements conflictuels provenant d'une erreur de détection d'un bruit de l'environnement comme une parole d'un participant. Enfin, sur la base des travaux de Heldner et Edlund (2010) sur le tour de parole, nous considérons un seuil d'une seconde comme le seuil limite entre une transition fluide et une transition considérée comme comportant un silence gênant.

Pour l'analyse des recouvrements nous avons repris les recouvrements extraits automatiquement par l'algorithme de détection des tours de parole puis à partir de cette extraction automatique nous avons sélectionné manuellement les recouvrements à analyser de sorte que :

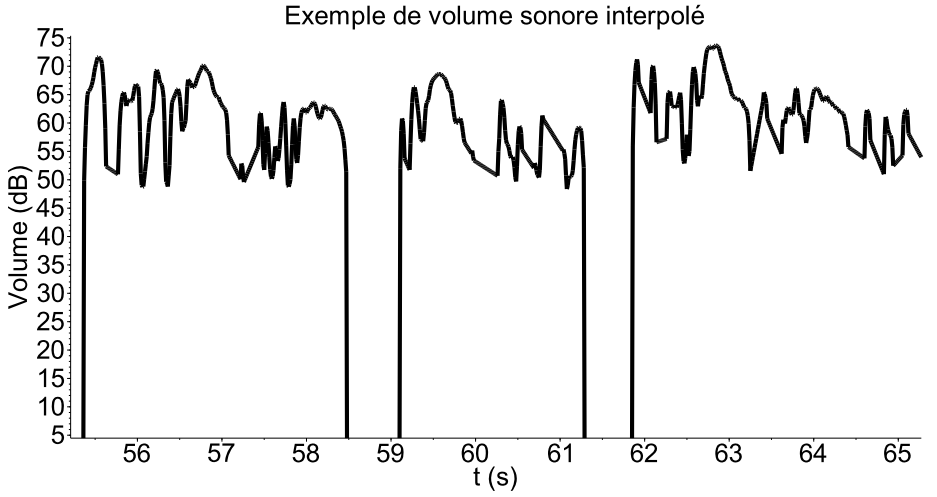
— le recouvrement ne soit pas coopératif : il ne doit pas être un *backchannel* produit en même temps que l'énoncé du locuteur, ni une complétion collaborative, ni une production chorale,



(a) Volume sonore tel qu'extrait par Praat.



(b) Volume sonore filtré selon les segments voiced/unvoiced renseignés par les valeurs de hauteur de voix.



(c) Volume sonore interpolé pour ne pas tenir compte des silences de moins de 400 ms.

FIGURE 9.3 – Exemple de traitement d'une portion de courbe du volume sonore. La figure 9.3a montre le profil de volume sonore tel qu'extrait par Praat, la figure 9.3b montre les valeurs de volume sonore filtrés en tenant compte de la distinction entre les segments parlés et non parlés, et la figure 9.3c montre le même profil que précédemment mais en interpolant les valeurs de volume sonore pour des micro-pauses de moins de 400 ms.

— le recouvrement détecté ne doit pas contenir de rires ou autres bruits parasites (souffle, bruit provenant de l'environnement).

Durées de transition

Pour mesurer les transitions de tour, nous nous sommes basés sur l'extraction automatique des tours de parole telle que présentée dans la section précédente. Ces transitions de tour peuvent, soit être un moment de silence, inférieur ou supérieur à 1 seconde, soit être des transitions avec un léger recouvrement non compétitif, soit être des transitions constituées d'une interruption de l'auditeur précédent. Pour chaque transition, la durée de recouvrement ou de silence est mesurée. Lorsque la transition comporte un recouvrement la valeur extraite est une valeur négative dont la valeur absolue est égale à la durée de recouvrement, lorsque la transition comporte un silence, la valeur extraite est une valeur positive dont la valeur absolue est égale à la durée de silence.

Nous avons analysé au total 411 transitions de parole provenant des 13 passations. La répartition des durées de transition est montrée sur la figure 9.4. On observe une répartition des durées de transition asymétrique, comportant plus de valeurs positives que négatives. Les durées vont de 2 secondes de recouvrement à 7 secondes de silence. La valeur médiane de durée de silence inter-transition est de 470 ms. Cette valeur médiane est proche de la valeur médiane de transition de tour de 451 ms trouvée par Campione et Véronis (2002) pour la langue française.

Intéressons-nous maintenant à la différence entre les durées de transition dans la condition où les participants se voient et dans la condition où les participants ne se voient pas. La différence entre les durées de transition est montrée sur la figure 9.5. En accord avec notre hypothèse 2 sur la dégradation de la coordination entre les participants, nous aurions pu penser que le nombre de recouvrements et de silences longs était significativement plus grand pour la condition sans présence de signaux non-vocaux par rapport à la condition avec présence de signaux non-vocaux tel que présenté par l'hypothèse 2. Nous constatons au contraire une tendance inverse ici : la condition où les participants se voient comporte plus de recouvrements et la durée moyenne de silence dans cette condition est plus grande que la condition où les participants ne se voient pas. La répartition des données ne suivant pas une loi normale et les mesures étant répétées, nous avons effectué un test de Mann Whitney pour savoir si cette tendance était significative. Les résultats ont montré que la différence dans les durées de transition entre les deux conditions n'était pas significative ($p=0.2$).

Analysons maintenant les durées des conflits de parole entre les participants. La distribution des conflits de parole pour les deux conditions est représentée sur la figure 9.6. Nous obtenons une valeur médiane de conflit de 700 ms dans la condition sans signaux non-vocaux et d'une seconde pour la condition avec signaux non-

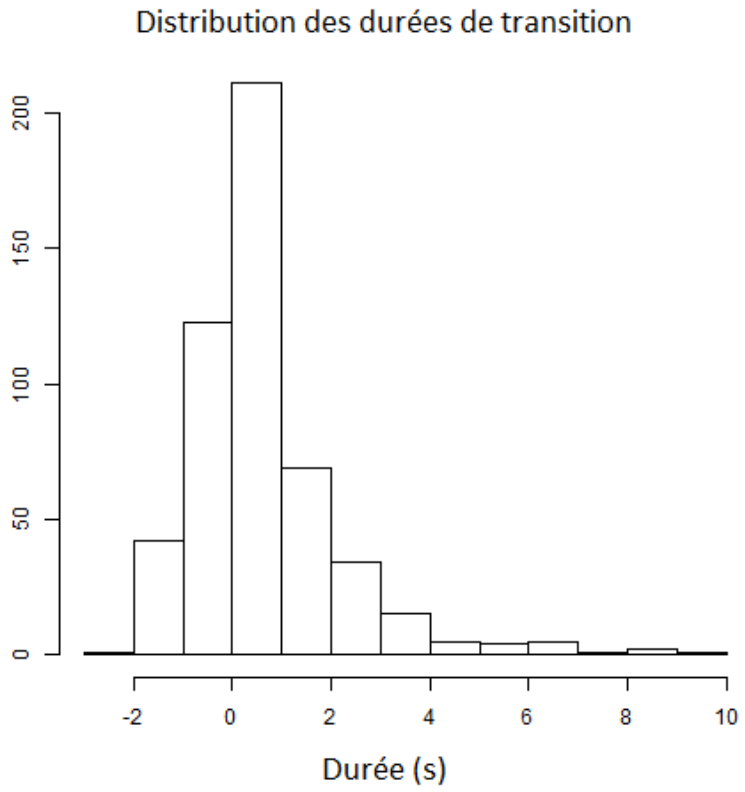


FIGURE 9.4 – Répartition des durées de transition (en seconde) observées dans l'expérimentation.

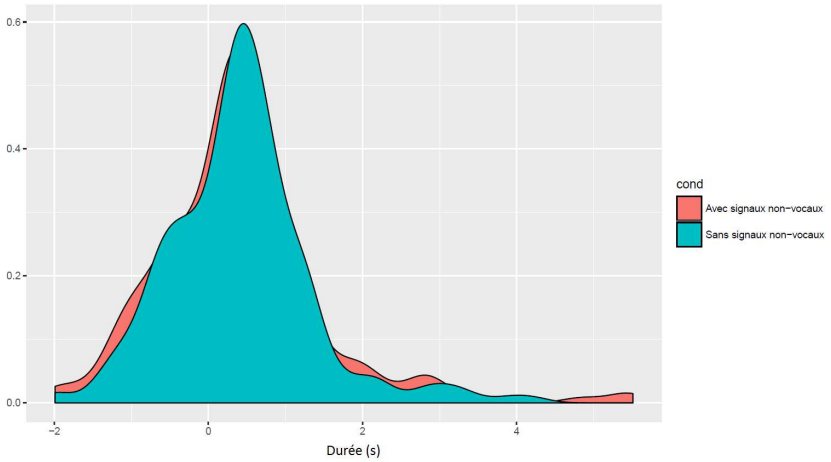


FIGURE 9.5 – Histogramme des durées de transition pour la condition avec (bleu) et sans signaux non-vocaux (rouge).

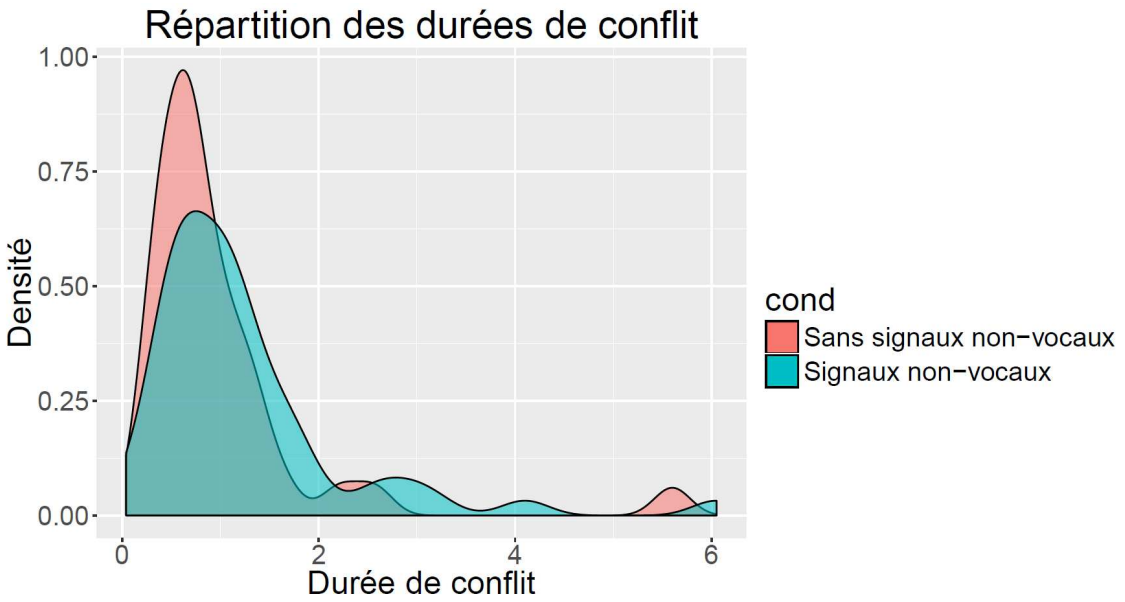


FIGURE 9.6 – Répartition des durées de conflit lors des échanges de parole entre les participants.

vocaux. Les données n'étant pas normales, nous avons effectué un test de Mann Whitney pour évaluer le caractère significatif de la différence entre les conditions. Les résultats du test ont conclu à une différence non significative entre les deux conditions (p=0.07).

Variation des signaux prosodiques

Nous avons mesuré les variations de volume sonore et de hauteur de voix du locuteur courant en fin de tour et des locuteur et auditeur courant lors de l'occurrence d'un conflit de parole. Pour cette mesure nous avons systématiquement comparé la valeur de volume sonore et de hauteur de voix lors des situations de fin de tour et de conflit de parole aux valeurs moyennes de volume sonore et de hauteur de voix du participant correspondant sur la totalité de l'interaction.

Nous avons choisi de distinguer les occurrences de recouvrement suivies d'une reprise de la parole par le locuteur précédant le conflit des occurrences de recouvrement suivies d'une prise de parole de l'auditeur précédent. Au total, les recouvrements suivis d'une reprise de parole du locuteur précédent représentent 69 % de la totalité des recouvrements extraits, les recouvrements suivis d'une prise de parole du locuteur suivant représentent 25 % de ces recouvrements, 6 % des recouvrements sont suivis d'un moment de silence. Ces dernières situations ne sont pas traitées ici. La distribution des transitions n'étant pas normale, des tests de Mann Whitney ont été utilisés pour comparer systématiquement la différence entre les conditions. Les résultats pour le locuteur courant sont présentés dans le tableau 9.2 et pour l'auditeur courant dans le tableau 9.3. Ces deux tableaux renseignent l'écart entre la moyenne du volume sonore et de la hauteur de voix lors de l'occurrence d'un

Le locuteur courant garde le tour			
Présence des signaux non-vocaux		Absence des signaux non-vocaux	
Volume sonore	Hauteur de voix	Volume sonore	Hauteur de voix
+ 5.37 dB ($p < 0.001^{***}$)	n.s. ($p = 0.491$)	+7.4dB($p = 0.041^{*}$)	n.s. ($p = 0.325$)
Le locuteur courant laisse le tour			
Présence des signaux non-vocaux		Absence des signaux non-vocaux	
Volume sonore	Hauteur de voix	Volume sonore	Hauteur de voix
+ 5.3 dB ($p = 0.006^{**}$)	-7.6 Hz ($p = 0.004^{**}$)	+5.9dB ($p = 0.012^{*}$)	-12 Hz ($p = 0.005^{**}$)

Table 9.2 – Variations des valeurs de volume sonore et de hauteur de voix du locuteur courant dans le cas de la résolution de conflit.

L'auditeur précédent abandonne sa prise de tour			
Présence des signaux non-vocaux		Absence des signaux non-vocaux	
Volume sonore	Hauteur de voix	Volume sonore	Hauteur de voix
+ 5.6 dB($p < 0.001^{***}$)	-23 Hz ($p = 0.06$)	+ 6.2 dB ($p = 0.01^{**}$)	n.s. ($p = 1$)
L'auditeur précédent prend le tour			
Présence des signaux non-vocaux		Absence des signaux non-vocaux	
Volume sonore	Hauteur de voix	Volume sonore	Hauteur de voix
+ 8.84 dB($p < 0.001^{***}$)	n.s. ($p = 0.465$)	+10.9 dB($p = 0.012^{*}$)	n.s. ($p = 0.29$)

Table 9.3 – Variations des valeurs de volume sonore et de hauteur de voix de l'auditeur précédent dans le cas de la résolution de conflit.

recouvrement conflictuel et la valeur moyenne du volume sonore et de la hauteur de voix sur toute l'interaction. Une valeur positive met en évidence un participant augmentant la valeur de ses signaux tandis qu'une valeur négative met en évidence un participant diminuant la valeur de ses signaux. Le volume sonore et la hauteur de voix sont renseignés dans ces tableaux pour les cas où le locuteur courant garde la parole (renommé par souci de clarté dans le tableau 9.3 « l'auditeur précédent abandonne sa prise de tour »), et les cas où le locuteur courant laisse la parole à l'auditeur précédent (renommé « l'auditeur précédent prend le tour » dans le tableau 9.3). Ces valeurs sont de plus distinguées pour la condition où les participants se voient (présence de signaux non-vocaux) et la condition où les participants ne se voient pas.

Les résultats montrent une augmentation du volume sonore lorsque les participants se trouvent en situation de recouvrement conflictuel. Ce résultat n'est pas surprenant puisque l'on retrouve ici les résultats trouvés par d'autres auteurs (Schegloff, 2000; Kurtić *et al.*, 2013). Lorsque l'on compare les valeurs de volume sonore entre les deux conditions, il semble que la valeur de volume sonore soit accentuée lorsque les participants ne se voient pas, en comparaison de la condition où les participants se voient. Cette constatation semble étayer notre hypothèse 1 : il semble

Condition	Volume Sonore	Hauteur de voix
Présence des signaux non-vocaux	-3.72 dB ($p < 0.001^{***}$)	-13.3 Hz ($p < 0.001^{***}$)
Absence des signaux non-vocaux	-3.64 dB ($p < 0.001^{***}$)	-12.7 Hz ($p < 0.001^{***}$)

TABLE 9.4 – Comparaison des valeurs de volume sonore et de hauteur de voix pour la fin de tour du locuteur courant.

bien y avoir une différence dans la manière dont les participants s'échangent leurs signaux selon l'absence ou non des signaux fournis par les participants. L'application de tests de Mann Whitney confirme cette tendance à l'augmentation des valeurs de volume sonore lorsque les participants ne se voient pas, laissant supposer que les participants accentuent bien les signaux à disposition lorsqu'ils ne se voient pas. Contrairement à ce qui avait été avancé par Schegloff (2000) et Kurtić *et al.* (2013), nous n'avons pas constaté de variation positive significative de la hauteur de voix lors de moments de conflictuels. La hauteur de voix varie négativement, par contre, lorsque les participants abandonnent leur tour (pour le locuteur) ou se ravisent de leur tentative de prise de tour (pour l'auditeur), cette variation ayant sans doute plus à voir avec des signaux de fins de tour, que des signaux spécifiques de résolution de conflit.

Pour l'analyse de la variation des signaux dans les fins de tour, nous avons sélectionné des fins de tour précédant des transitions comportant un moment de silence positif et inférieur à 1 s. Ces transitions ont été extraites à partir de l'annotation automatique des tours de parole que nous avons effectuée. Une étape de sélection manuelle a ensuite été réalisée pour éliminer :

— les transitions où le locuteur suivant ne fait que produire un *backchannel*,
— les transitions comportant des rires ou des bruits parasites (souffles dans le micro, ou bruit provenant du réajustement du micro).

Nous montrons dans le tableau 9.4 les valeurs de baisse de volume sonore et de hauteur de voix. Ces valeurs ont été mesurées de la même manière que pour les situations de recouvrement, en soustrayant la valeur du signal avec la valeur médiane de signal du participant sur la totalité de l'interaction.

Les observations confirment les résultats observés par la majorité des auteurs de la littérature : on observe bien une baisse de volume sonore et de hauteur de voix en fin de tour. Nous avons voulu vérifier par un test de Mann Whitney si la baisse de ces signaux prosodiques était significativement différente selon la condition. Nous n'avons pas trouvé de différences significatives.

Discussion

Nous avons montré la présence d'une variation de volume sonore et de hauteur de voix lors des fins de tour et une variation de volume sonore lors des conflits. Lors des moments de conflit, que le locuteur précédent reprenne le tour ou que

l'auditeur précédent prenne le tour, nous avons observé une accentuation significative du volume sonore entre les conditions (+10 dB pour l'auditeur). Cependant nous n'avons pas trouvé de variation significative de la hauteur de voix lors de moment de conflit. Ce résultat est en contradiction avec des études récentes sur le tour de parole (Schegloff, 2000; Kurtić *et al.*, 2013; Chowdhury *et al.*, 2015). Néanmoins, ces études récentes ont été effectuées pour les langues anglaise et italienne et les variations de signaux pourraient être différentes pour la langue française. Le degré de connaissance des participants peut aussi être mis en avant. L'augmentation de la hauteur de voix pourrait être un marqueur de dominance dans la conversation. Or le dispositif expérimental fait ici interagir deux participants ne se connaissant pas.

Si l'on s'intéresse maintenant aux variations des durées de transition entre la condition où les participants se voient et la condition où les participants ne se voient pas, on n'observe pas de variation significative de durée de transition et de durée de conflit, invalidant l'hypothèse HO 2. Le fait que nous n'ayons pas trouvé de différences significatives dans les durées de conflit entre les deux conditions pourrait venir du nombre de conflits analysés (51 pour la condition avec signaux non-vocaux et 36 pour la condition sans signaux non-vocaux). Le degré de significativité trouvé (p=0.07) montre néanmoins une tendance vers une différence dans la durée des conflits. La variation des valeurs de durée de conflit entre les deux conditions est de plus grande : 700 ms pour la condition où les participants ne se voient pas contre 1 seconde pour la condition où les participants se voient. Nous constatons donc, au vu des résultats, qu'une coordination est toujours présente lorsque les participants ne se voient pas, confirmant l'hypothèse HO1. Il n'existe néanmoins pas de dégradation de la coordination des participants. De plus, la présence ou l'absence de signaux non-vocaux n'impactent pas non plus sur les variations de signaux en fin de tour, la décroissance des signaux se fait au même moment et de la même manière. Une adaptation existe cependant pour les moments de conflit, les participants tendent à accentuer leur volume sonore quand ils ne se voient pas. La différence dans la variation du volume sonore dans les situations de conflit entre les deux conditions est en accord avec les résultats observés lors des simulations agent-agent, et pourrait être un indicateur du couplage dans la production et l'interprétation du volume sonore des deux participants. L'absence de différence de variation dans le cadre de fins de tour semble mettre en cause l'hypothèse de couplage dans ces situations précises. Néanmoins, cela pourrait indiquer une contribution faible des variations des signaux non-vocaux dans la transmission de la parole comme l'affirme Sellen (1995).

9.2 Paramétrage du modèle

À partir des observations des interactions humaines, une implémentation du modèle a été réalisée. Nous avons d'abord défini les fonctions $f(\gamma, m)$ des équations

Auditeur	Fonction
Volume	$f(\gamma, m) = 0.5 \times S(2 \times (2 \times m + \gamma - 0.8)) + (m - 0.5) \times S(-2 \times \gamma) \times S(2 \times (m + \gamma - 0.8))$
Hauteur de voix	$f(\gamma, m) = 0.5 \times S(2 \times m + \gamma - 0.9)$
Locuteur	Fonction
Volume	$f(\gamma, m) = 0.5 \times S(10 \times (-m - \gamma)) \times S(10 \times m) + 0.5 \times S(-10 \times m) - 0.5 \times m \times S(10 \times \gamma) \times S(-10 \times m)$
Hauteur de voix	$f(\gamma, m) = 0.5 \times (S(-m - \gamma) \times S(m) + S(-m))$

TABLE 9.5 – Équations de contrôle des signaux prosodiques des agents.

de contrôle des signaux (voir équation 6.3 page 112) sur la base des variations de volume sonore et de hauteur de voix observées dans les interactions humaines. Pour la conception de ces équations, nous avons suivi la même méthodologie que celle présentée section 7.1 chapitre 7. Les équations déduites des interactions humaines sont présentées dans le tableau 9.5.

La fonction S désigne ici la fonction sigmoïde présentée dans le chapitre 6, page 113. m désigne la motivation du participant, γ désigne le niveau de certitude du participant. Pour l'implémentation du modèle, nous avons défini les grandeurs de volume sonore et de hauteur de voix de sorte que :

— 0.5 indique une valeur moyenne de signal,
— 1.0 indique une valeur maximale de signal,
— 0.0 indique pour le volume sonore un arrêt de la parole, et une valeur minimale pour la hauteur de voix.

Nous n'avons pas fixé à quoi correspond concrètement les différentes valeurs des signaux, la correspondance entre les valeurs théoriques provenant de notre modèle et la prosodie réelle de l'agent dépend de l'environnement d'interaction avec l'utilisateur ou encore de la voix de synthèse.

La figure 9.7 illustre comment les équations de contrôle du volume sonore assurent le lien entre la motivation propre de l'agent à parler m et la certitude dans la motivation de l'interlocuteur γ. Pour le locuteur courant, nous avons déterminé les équations de sorte que lorsque sa valeur de motivation est inférieure à 0 et le niveau de certitude est supérieure à 0, tout en restant inférieur en valeur absolue à la motivation à garder le rôle, son attracteur augmente vers la valeur 1 (le volume sonore de l'agent converge vers 1). Lorsque la valeur de certitude dépasse la valeur de motivation de l'agent, le volume sonore de ce dernier décroit vers 0, simulant un participant laissant le tour à son partenaire. De la même manière, lorsque la motivation est positive, le locuteur peut initier un changement de rôle même s'il pense que l'autre agent ne veut pas changer de rôle (certitude négative concernant le changement de rôle de l'autre agent). Dans ce cas de figure, moins l'agent aura une motivation grande à laisser le tour, plus il faudra une certitude positive pour qu'il initie un abandon de tour. Pour l'auditeur courant, l'augmentation du volume sonore

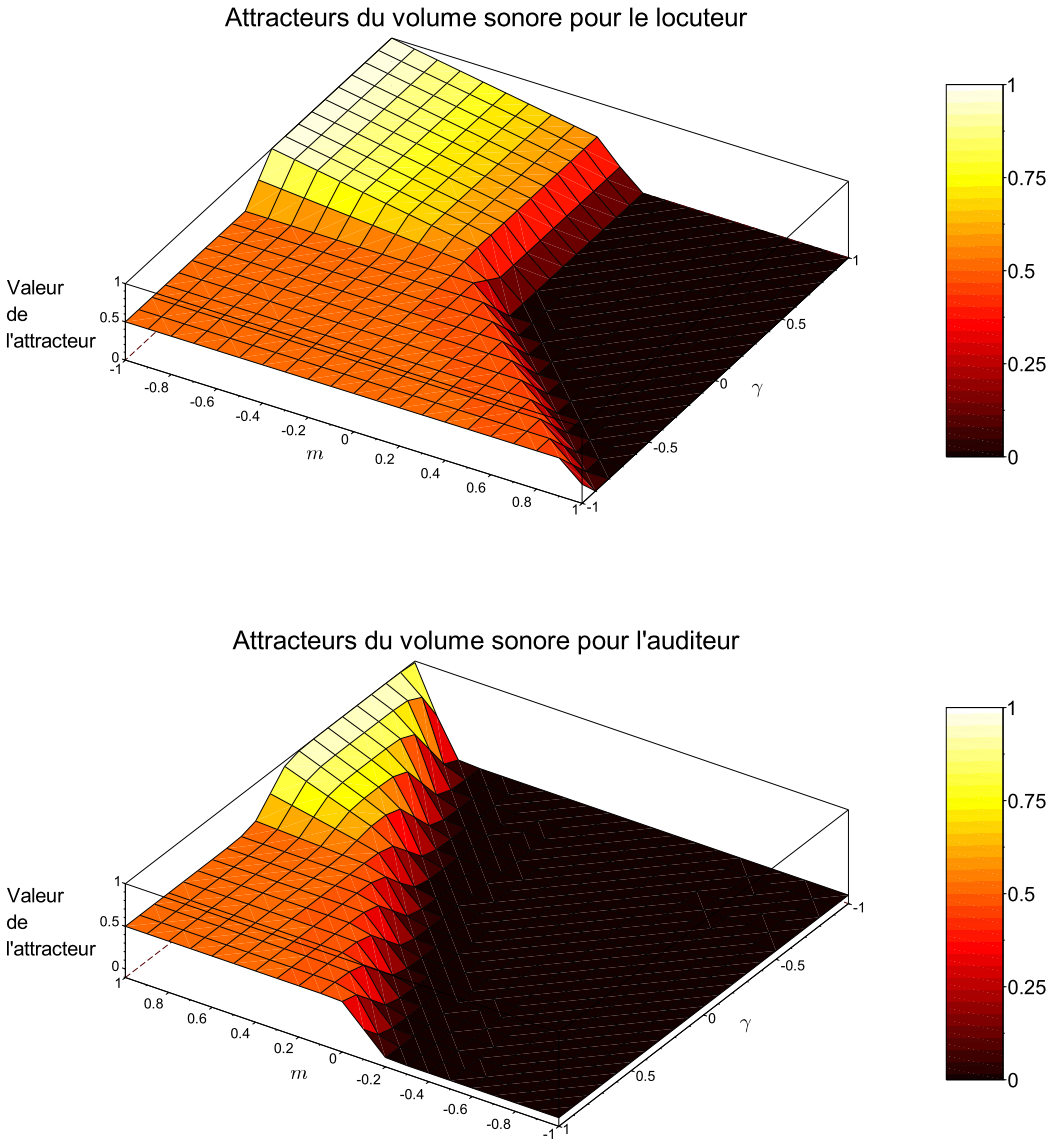


FIGURE 9.7 – Valeurs vers lesquelles le volume sonore converge pour le locuteur courant (haut) et l'auditeur courant (bas) en fonction de m et γ.

est aussi dépendante de la valeur de motivation. Lorsque $m = 0$, il faut une valeur de certitude de 1 pour que l'agent augmente son volume sonore à 0.5, tandis que l'agent n'attend pas de valeur de certitude positive. Lorsque sa motivation est égale à 1, il prend la parole en augmentant son niveau sonore à 1 si la valeur de certitude est négative et à 0.5 si la valeur de certitude est positive. Les fonctions définissant les attracteurs de la hauteur de voix ont été définies pour avoir des valeurs similaires aux valeurs de volume sonore, en supprimant néanmoins les variations de hauteur de voix lors de situations conflictuelles en accord avec les données expérimentales.

Nous avons défini les valeurs d'amortissement et de raideur de chaque équation de sorte que les variations de volume sonore et de hauteur de voix en fin de tour correspondent à ce que l'on a observé. Ainsi, pour avoir une décroissance de signal similaire à ce que nous avons observé dans les interactions humaines (décroissance à partir de 300 ms avant la fin de tour), nous avons fixé l'amortissement et la raideur de toutes les équations de contrôle du signal pour le locuteur et l'auditeur à $b = 26.0$ et $k_g = 160.0$.

L'implémentation du module de prise de décision consiste en l'instanciation de chaque fonction d'accumulation α_j présentée dans l'équation 6.2, page 103. Ces fonctions ont été approximées par des polynômes, représentant des approximateurs universels de fonction. Ces polynômes se présentent sous la forme montrée par l'équation 9.1 :

$$y = b_0 + b_1 s_j + b_2 \dot{s_j} \tag{9.1}$$

Cette équation propose à la fois l'analyse de la dérivée de volume sonore et de hauteur de voix avec la présence d'un terme du second degré, l'analyse de la prosodie de l'autre participant est ainsi plus riche que ce qui a été présenté chapitre 7. Les paramètres de l'équation ont été calculés de sorte que, dépendant des motivations des agents, nous observons des transitions fluides et des recouvrements en cohérence avec les motivations de chaque agent. Un conflit doit être observé lorsque les deux participants ont des motivations m de signes différents. Si la motivation du locuteur $m_l > 0$ et celle de l'auditeur $m_a < 0$, un moment de silence est observé et si $m_l < 0$ et $m_a > 0$, un recouvrement est observé.

Un algorithme de recuit simulé a été utilisé pour déterminer les paramètres des équations de perception afin de faire correspondre aux interactions agent-agent les durées de conflit et de transition observées dans les interactions humaines. Pour l'apprentissage de cette implémentation, le participant ayant le tour à la fin de la simulation est systématiquement comparé avec le participant étant attendu comme possesseur du tour. Nous avons défini le possesseur courant du tour de sorte que dans une situation de conflit, c'est le participant ayant la valeur de motivation la plus grande qui deviendra le locuteur suivant. Lors d'une situation de changement de tour non conflictuelle, nous avons ajouté comme critère d'apprentissage la reproduction des durées de transitions observées dans les interactions humaines. Nous

avons ainsi cherché à faire varier la durée de transition d'un léger recouvrement de parole de 300 ms (représentant la durée de recouvrement non conflictuel la plus importante) lorsque les deux participants ont une motivation de changer de rôle de 1.0 et un moment de silence de 900 ms (les durées de silence après 900 ms étant moins fréquentes) lorsque la motivation du locuteur est égale à 1.0 et la motivation de l'auditeur à prendre le tour est égale à -1.0. Un critère similaire a été fixé pour les durées de conflits allant de 700 ms pour la motivation de l'auditeur à 0.6 et la motivation du locuteur à -1.0, jusqu'à 1.2 s pour la motivation de l'auditeur à 1.0 et la motivation du locuteur à -1.0.

9.3 Simulation agent-agent

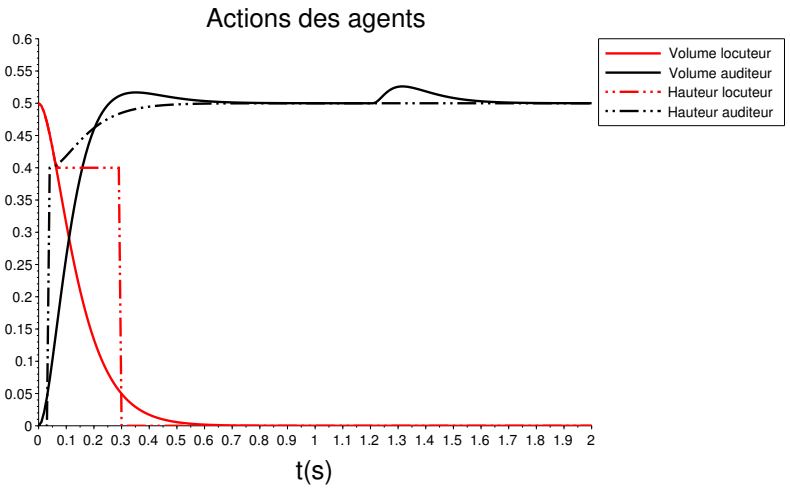
Nous présentons dans cette section les résultats de simulations agent-agent montrant la capacité des agents à reproduire les durées de transition et les variations de signaux observées dans les interactions humaines. Les résultats ont montré une répartition des durées de transition entre de légers recouvrements de l'ordre de 300 ms (motivation des deux agents à 1.0) et des moments de silence de 900 ms (motivation du locuteur à 1.0 et motivation de l'auditeur à -1.0). La figure 9.8 montre les deux situations extrêmes et les valeurs associées.

De plus, en accord avec les observations, le locuteur courant commence à décroitre son volume sonore et sa hauteur de voix 300 ms avant la fin du tour.

L'analyse des conflits montre des durées de recouvrement entre 700 ms (motivation de l'auditeur à 0.6 et motivation du locuteur à -1.0) et 1.2 s (motivation de l'auditeur à 1.0 et motivation du locuteur à -1.0). Les différents cas sont exposés dans la figure 9.9. Dans les deux scénarios présentés sur cette figure, l'auditeur courant tente d'interrompre le locuteur suivant. Ce dernier, percevant la tentative de prise de tour de son partenaire, augmente son volume sonore. Selon la valeur de motivation de l'auditeur courant, le conflit résulte en une tentative avortée de l'auditeur courant (figure du haut) ou une interruption (figure du bas).

Enfin, nous avons évalué la capacité de notre modèle à agir de manière cohérente sur plusieurs tours. La figure 9.10 montre un scénario où la motivation des participants varie selon une machine à états que nous avons définie. Les courbes en rouge et bleu au centre de la figure représentent la variation de volume sonore pour chacun des agents. En dessous, les courbes rouges et bleues représentent l'évolution temporelle des motivations m des deux agents. Superposées à ces courbes sont représentées les valeurs de certitude γ des deux agents.

Plusieurs situations émergent de ce scénario. La première situation correspond à une transition fluide entre les deux participants. Dans la seconde situation, l'agent représenté par la courbe bleue laisse la parole à l'agent représenté par la courbe rouge. Constatant que son interlocuteur ne prend pas la parole, le locuteur précé-



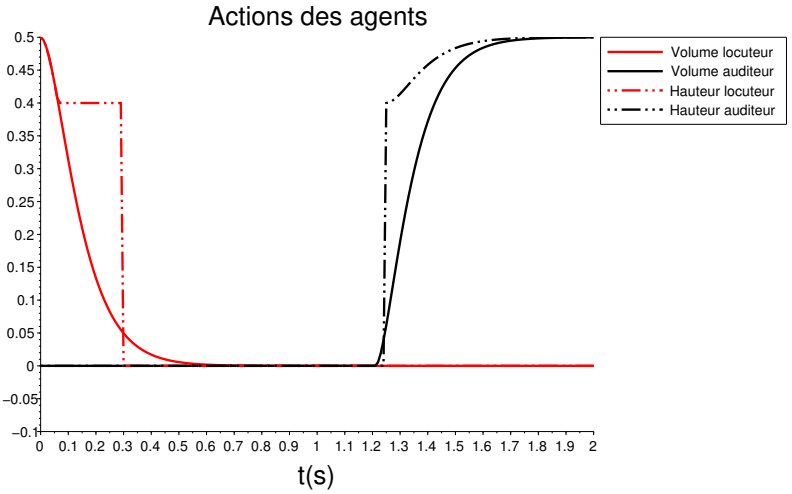


FIGURE 9.8 – Deux transitions fluides produites par le modèle. La figure de gauche montre un recouvrement léger de 250 ms et la figure de droite un moment de silence de 900 ms. Courbe rouge : locuteur précédent ; courbe bleue : locuteur suivant.

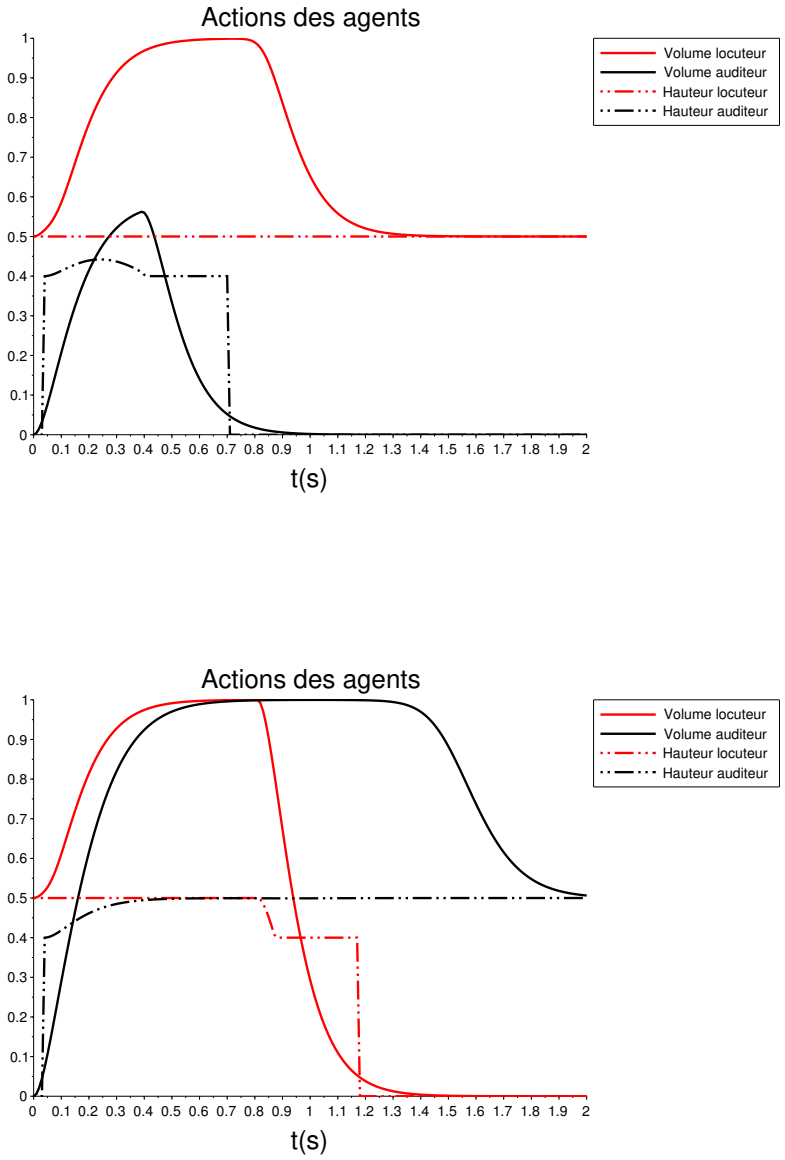


FIGURE 9.9 – Deux recouvrements compétitifs. La figure du haut montre un recouvrement de 700 ms et la figure du bas un recouvrement de 1.2 s.

dent reprend la parole, c'est à ce moment que l'auditeur précédent prend la parole, résultant en un recouvrement simultané non voulu, résolu par le locuteur précédent laissant la parole. Ces situations de recouvrement non voulues ont été couramment observés dans notre corpus, et il ne sera pas rare que l'agent doive s'adapter très rapidement à ces situations dans le cas d'interactions utilisateur-agent. Ici, l'observation montre un agent capable de s'arrêter de parler rapidement après la prise de parole de l'agent en rouge. La dernière transition montre un conflit entre les deux agents. Les valeurs de motivation ont été fixées à -1.0 pour l'agent en rouge, indiquant qu'il souhaite fortement garder la parole, et à 0.8 pour l'agent en bleu indiquant qu'il souhaite prendre la parole. Plutôt que de montrer une situation classique où les

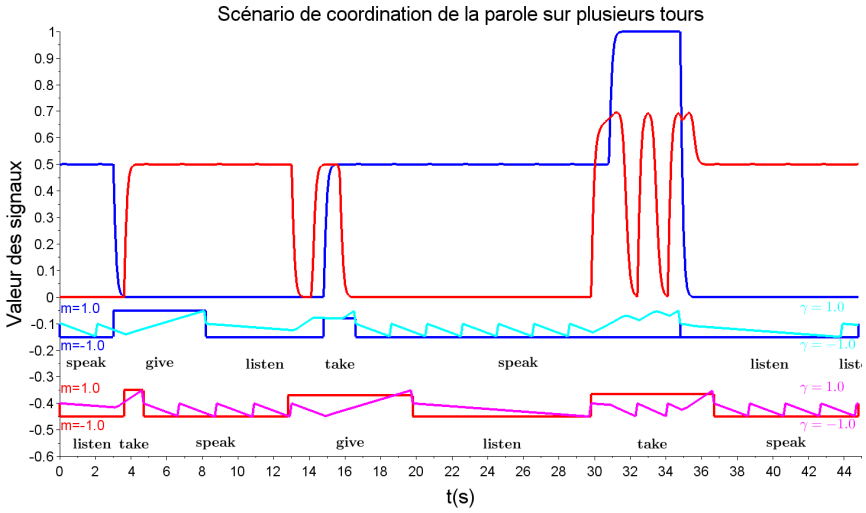


FIGURE 9.10 – Exemple d'un scénario sur plusieurs tours. En haut, en bleu, le volume sonore du locuteur, en rouge le volume sonore de l'auditeur. En-dessous, en bleu, la motivation du locuteur, en rouge, la motivation de l'auditeur, en cyan, la valeur d'accumulation du locuteur, en magenta, la valeur d'accumulation de l'auditeur.

participants gardent indéfiniment un niveau sonore élevé tant que les valeurs de motivations ne changent pas, la figure montre un auditeur précédant effectuant d'abord deux tentatives avortées de prendre la parole puis réussissant à prendre la parole à l'agent en rouge dans sa troisième tentative.

9.4 Conclusion

Nous avons présenté dans ce chapitre une validation et une calibration de notre modèle à partir de l'observation d'interactions humaines. Nous avons proposé un protocole expérimental où deux participants sont chargés de négocier pour accomplir un objectif commun. Pour valider l'hypothèse de couplage entre agents, nous avons observé les variations de durée de transition entre une condition où les agents ont accès aux signaux non-vocaux et une condition où les participants n'ont pas accès aux signaux non-vocaux de leur partenaire. Nous n'avons pas observé de variations de durées de transition ou de conflit lors des transitions de tour ni de variations de signaux en fin de tour, montrant que l'absence de signaux visuels ne semble pas affecter la transmission de tour. Néanmoins, une adaptation existe lors des moments de conflit entre les deux participants en accord avec notre hypothèse de couplage. Nous avons ensuite présenté une implémentation de notre modèle à partir des données récoltées de l'expérimentation et nous avons montré la capacité de deux agents pilotés par notre modèle à se coordonner en reproduisant les variations de signaux observées dans les interactions humaines. Nous présentons dans le chapitre suivant

la manière dont nous avons implémenté notre modèle dans l'architecture d'agent présenté dans le chapitre 8 et l'évaluation utilisée pour analyser le ressenti de l'utilisateur en interaction avec l'agent.

Chapitre 10

Interaction agent-utilisateur

Le chapitre 9 a montré la capacité de notre modèle à simuler l'interaction entre deux agents qui reproduit certaines propriétés des interactions conversationnelles entre humains. Notre objectif est maintenant de valider la capacité de notre modèle à assurer la coordination des tours de parole entre un agent et un humain. Nous présentons d'abord en section 10.1 l'implémentation complète du modèle dans l'architecture BeAware et les choix techniques que nous avons adoptés pour obtenir une réalisation opérationnelle de notre proposition. Ensuite, en section 10.2, nous montrons la capacité de notre solution à générer le comportement d'un agent capable d'interagir en temps réel avec un utilisateur, la coordination étant ici basée uniquement sur les signaux prosodiques. Enfin, en section 10.3 nous rendons compte des résultats de l'étude que nous avons réalisée pour étudier l'influence de la stratégie de gestion des tours de paroles sur le ressenti de l'utilisateur. Par rapport aux travaux antérieurs sur la gestion du tour de parole, l'originalité de cette évaluation réside dans la pris en compte, non seulement des capacités du modèle à assurer une bonne coordination des tours de parole (par exemple, la capacité à l'agent à réagir à une intervention coopérative ou compétitive, la capacité à discriminer une fin de tour d'une pause intra-tour), mais aussi dans le ressenti rapporté par les utilisateurs vis-à-vis de leur interaction avec l'agent.

10.1 Implémentation des équations du modèle

L'implémentation des équations du modèle dans l'architecture d'agent proposée dans le chapitre 8 pose plusieurs problématiques à résoudre.

1. De quelle manière sont implémentées la perception du comportement de l'utilisateur et le contrôle des actions de l'agent ? Implémente-t-on l'accumulation partielle des signaux de l'agent et le calcul de la certitude en un sous-module ou en plusieurs sous-modules s'exécutant en parallèle ?
2. Où sont implémentés les sous-modules responsables de la perception du comportement de l'agent et du contrôle de l'agent ?

3. Comment établissons-nous le lien entre la gestion du tour de parole et l'interprétation et le contrôle des énoncés produits par l'agent ?

Nous présentons dans la suite la manière dont nous avons résolu ces problématiques.

10.1.1 Perception du comportement de l'utilisateur

Dans notre modèle la perception du comportement de l'utilisateur peut être décomposée en plusieurs processus distincts : le calcul de chaque valeur d'accumulation partielle (fonctions α_j de l'équation 6.2, page 103), le calcul de la valeur d'accumulation par l'équation 6.2 à partir des valeurs d'accumulation partielles, le calcul de la variable de certitude par l'équation 6.1, page 103 et l'action prise une fois un seuil de certitude positif franchi. Dans notre modèle, l'exécution de chaque processus s'effectue indépendamment des autres. Le calcul de la valeur de certitude ne nécessite pas la mise à jour de la valeur d'accumulation et le calcul de la valeur d'accumulation ne nécessite pas que le module calculant la valeur de certitude ait mis à jour la variable de certitude par exemple. Au contraire, chaque processus ne se préoccupe que de récupérer l'information disponible au moment de son exécution. Le processus d'accumulation peut être lui-même décomposé en une somme d'accumulations partielles et la relation entre le calcul de la valeur d'accumulation utilisée par le processus de perception et le calcul des valeurs d'accumulations partielles suivent la même logique. Nous proposons, suivant l'approche modulaire d'Ymir (Thórisson, 1999), de séparer ces différents processus en sous-modules indépendants. Nous aurons donc un sous-module de perception calculant la valeur de certitude en s'appuyant sur la valeur d'accumulation, un sous-module de calcul de la valeur d'accumulation qui s'appuie sur les résultats calculés par chaque sous-module d'accumulation partielle, et un sous-module se chargeant de changer l'état de l'agent lorsque la certitude aura franchi le seuil positif. Les processus s'occupant de l'action à réaliser lors du franchissement de chaque seuil surveillent l'état de la valeur de certitude et produisent une action quand cette valeur est supérieure au seuil de décision.

Déterminons maintenant dans quel module ces sous-modules seront exécutés. L'architecture présentée chapitre 8 laisse trois possibilités pour l'implémentation de ces sous-modules. Le module de captation s'occupe de la récupération et la transmission des informations provenant des capteurs de l'agent sans intégration multimodale de ces données sensorielles. Ce n'est donc pas dans ce module que nous implémenterons nos sous-modules. Il reste alors deux modules possibles pour l'implémentation des sous-modules. Le module d'interprétation du comportement se charge de la perception unimodale ou multimodale de comportements produits par l'utilisateur tels que les hochements de tête. Le module d'interprétation de la fonction communicative associe un état du dialogue à l'utilisateur courant. Dans une application de l'architecture ASAP effectuée par Kopp *et al.* (2014), chaque comportement de l'utilisateur lié au tour de parole est modélisé en états (« état » de

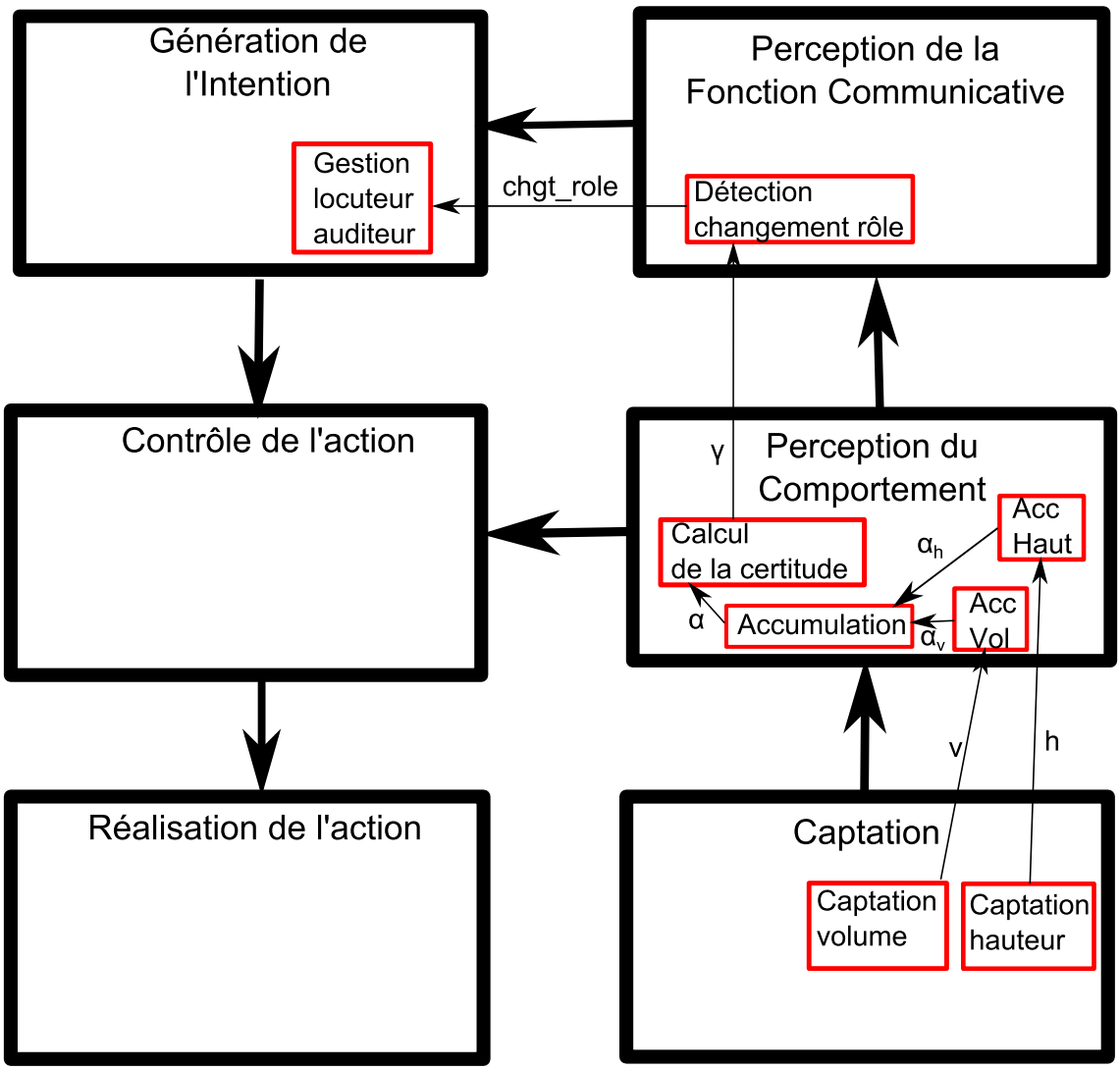


FIGURE 10.1 – Illustration de l'implémentation des modules de perception dans l'architecture.

vouloir prendre le tour, laisser le tour, garder le tour). Les processus percevant le comportement de l'agent sont alors implémentés dans le module de perception de la fonction communicative. Au contraire, notre modèle n'analyse pas les intentions de l'utilisateur mais perçoit directement le comportement de l'utilisateur de la même manière qu'il aurait perçu un hochement de tête de la part de son partenaire. En ce sens, nous proposons l'implémentation des modules d'accumulation et de calcul de la valeur de perception dans le module se chargeant de l'interprétation du comportement de l'utilisateur. La notification du franchissement d'un seuil de décision est réalisée dans la couche d'interprétation de la fonction communicative. Le sous-module correspondant notifie alors un gestionnaire de module qui se charge de changer l'état du dialogue (« locuteur » ou « auditeur »), d'activer les sous-modules de perception et de contrôle de l'action liés au nouvel état et de désactiver les sous-modules de perception et de contrôle de l'ancien état. Nous illustrons la répartition des sous-modules de perception du comportement de l'agent sur la figure 10.1.

10.1.2 Contrôle des actions

Dans notre modèle, chaque signal de l'agent est modulé par une équation distincte. Ces équations sont exécutées en parallèle, indépendamment des autres équations. Ce choix peut paraitre contradictoire au vu du couplage physique existant entre les variations de signaux prosodiques et le fait que l'agent est en train ou non de parler. En ce sens, les différentes variations prosodiques sont couplées au volume sonore de l'agent. Lorsque le volume sonore est nul, l'agent ne parle pas et moduler la hauteur de voix n'a pas de sens. Néanmoins, ces équations de contrôle définissent avant tout les paramètres de contrôle des réaliseurs de l'agent et ne correspondent pas à la valeur réelle de ces signaux. Nous autorisons ainsi les équations de contrôle à moduler leur sortie respective indépendamment des autres équations de contrôle de l'agent. C'est alors, selon notre vision, du ressort du réaliseur de l'agent de gérer le couplage physique existant entre les différents signaux de l'agent. De la même manière que pour la perception, nous définissons donc chaque action de l'agent dans des sous-modules séparés, indépendants l'un de l'autre.

Intéressons-nous maintenant à l'emplacement des différents sous-modules de contrôle des actions de l'agent. Deux modules sont susceptibles de contenir ces sous-modules : la gestion de l'intention et le contrôle de l'action. Dans l'exemple d'application de l'architecture ASAP à la gestion du tour de parole proposé par Kopp *et al.* (2014), les différents comportements possibles de l'agent vis-à-vis du tour de parole sont formulés sous forme d'intentions de l'agent. L'agent a ainsi l'intention de garder le tour, prendre le tour ou laisser le tour. Par exemple, l'agent peut générer une intention de garder le tour lorsqu'il reçoit un événement signalant l'intention de l'utilisateur de prendre le tour. Cette intention de garder le tour est alors transcrite directement par les sous-modules du module de contrôle de l'action en variation des signaux de l'agent. Dans notre modèle, le principe de couplage implique que la perception du comportement de l'agent influe directement sur le contrôle des actions de l'agent. Les variations d'action de l'agent ne sont pas issues uniquement d'une intention ou d'un comportement planifié par l'agent comme proposé par Kopp *et al.* (2014) mais sont influencées par, à la fois, sa motivation à parler ou non et la perception du comportement de l'utilisateur. La variable de motivation de notre modèle peut être considérée comme une intention de l'agent. En ce sens, bien que nous ne modélisons pas dans notre modèle la manière dont la valeur de motivation peut varier, nous considérons que les sous-modules responsables de la variation de la motivation de l'agent sont implémentés dans le module de génération de l'intention et la valeur de motivation est transmise depuis le module de génération de l'intention. Les sous-modules de contrôle des signaux de l'agent n'étant, eux, pas spécifiquement issus d'une intention de l'agent, ils sont donc implémentés dans le module de contrôle de l'action. Ces sous-modules ont alors en entrée la valeur de motivation provenant de la couche de génération de l'intention et la valeur de certitude provenant de la

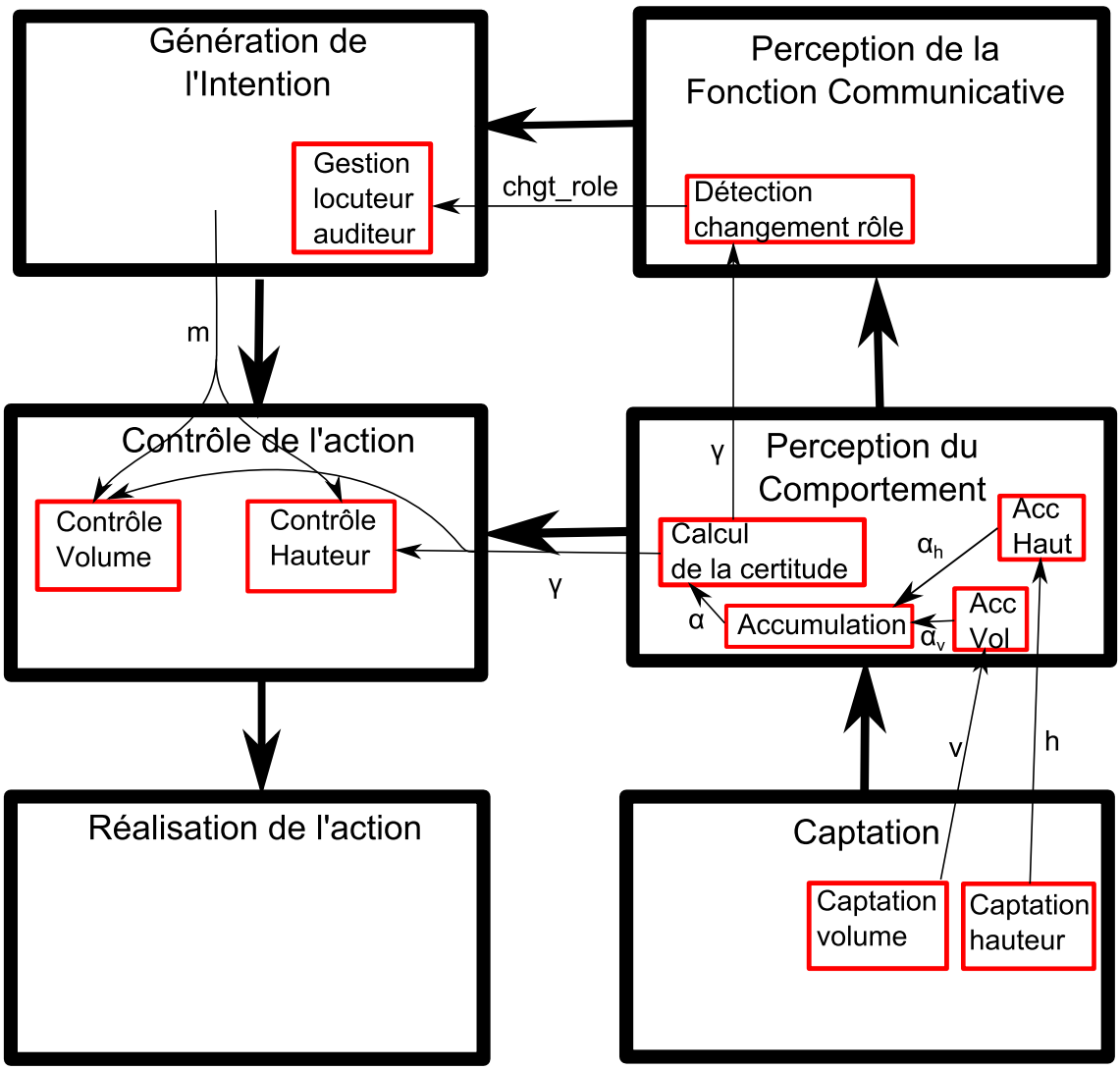


FIGURE 10.2 – Illustration de l'implémentation des sous-modules de perception et de contrôle de l'action dans l'architecture.

couche de perception du comportement. La figure 10.2 présente le même schéma que la figure 10.1 complété par les sous-modules de contrôle des actions.

10.1.3 Lien entre gestion du dialogue et gestion du tour de parole

Jusqu'à présent les simulations effectuées entre deux agents n'ont pas pris en compte le lien entre interprétation de l'énoncé, génération d'énoncé et gestion du tour de parole de la part de l'agent. Pour une interaction entre utilisateur et agent nous devons clarifier le lien entre l'interprétation du contenu de l'énoncé de l'utilisateur, la décision de produire un énoncé et le contrôle du tour de parole. Selon notre vision, la motivation à parler provient de différents facteurs dont les attitudes interpersonnelles et surtout le caractère important ou non de l'énoncé que l'agent a à produire. Si l'agent n'a rien à dire sa motivation sera ainsi négative, tandis que s'il a quelque chose à dire sa motivation sera positive. En ce sens la formulation de l'énoncé influe

directement sur la motivation à parler de l'agent. Le calcul de la motivation est réalisé dans un module indépendant du sous-module de planification de l'énoncé.

L'interprétation de l'énoncé de l'agent est quant à elle un processus habituellement traité de manière symbolique. Nous implémentons donc l'interprétation du contenu de l'énoncé de l'utilisateur dans le module de perception de la fonction communicative.

Lorsque l'agent a décidé de produire un énoncé, l'énoncé est-il immédiatement envoyé puis déclenché par la synthèse vocale ? Nous considérons que non. Une fois que l'agent a planifié un énoncé, le lancement de l'énoncé est géré par le réaliseur de l'agent. Lorsque l'agent planifie un énoncé, cet énoncé est transmis à un sous-module spécifique de la couche de contrôle de l'action. Ce sous-module évalue la contrainte associée à l'énoncé (IMMEDIATELY, START_IMMEDIATELY_AFTER par exemple). Si la contrainte IMMEDIATELY est définie, ce sous-module envoie l'énoncé au gestionnaire d'action qui se charge de transmettre la requête au réaliseur correspondant de l'agent. Si une contrainte START_IMMEDIATELY_AFTER avec un énoncé spécifié est reçu, le sous-module attendra le retour du réaliseur concernant la fin de l'énoncé spécifié pour le transmettre.

Le réaliseur chargé de produire les énoncés de l'agent fournit une interface entre l'architecture d'agent et le synthétiseur vocal. Il reçoit les requêtes formulées en BMLa envoyées par le gestionnaire d'action, et commande la synthèse vocale selon la requête reçue. Lorsque le réaliseur reçoit une commande de génération d'énoncé, il vérifie les valeurs de volume sonore et de hauteur de voix provenant des requêtes de l'agent. Si la valeur de volume sonore est supérieure à un seuil (défini actuellement à 0.2), il transmet l'énoncé à la synthèse vocale qui joue le flux audio. En parallèle de l'énoncé, il reçoit les requêtes de modulation de volume sonore et de hauteur de voix. Il établit alors la correspondance entre la hauteur de voix et le volume sonore, puis envoie la modification à la synthèse vocale qui se charge d'appliquer le changement à l'énoncé qui est en train d'être prononcé. Sur la figure 10.3 nous complétons le schéma 10.2 en ajoutant les modules chargés d'interpréter, de planifier et de produire les énoncés produits par l'agent.

10.1.4 Implémentation des modules

L'implémentation de l'architecture et des sous-modules a été réalisée en JAVA. Lors de cette implémentation nous avons fixé la fréquence d'exécution des modules de l'architecture à 100 Hz. L'implémentation des sous-modules de contrôle des actions et de calcul de la valeur de certitude a nécessité l'utilisation d'algorithmes de simulation numérique d'équations différentielles déterministes et stochastiques. Pour la simulation des équations de contrôle des actions nous avons implémenté la méthode de Runge Kutta 4. Pour la simulation de l'équation de perception du comportement, stochastique, nous avons utilisé la méthode d'Euler-Maruyama. La

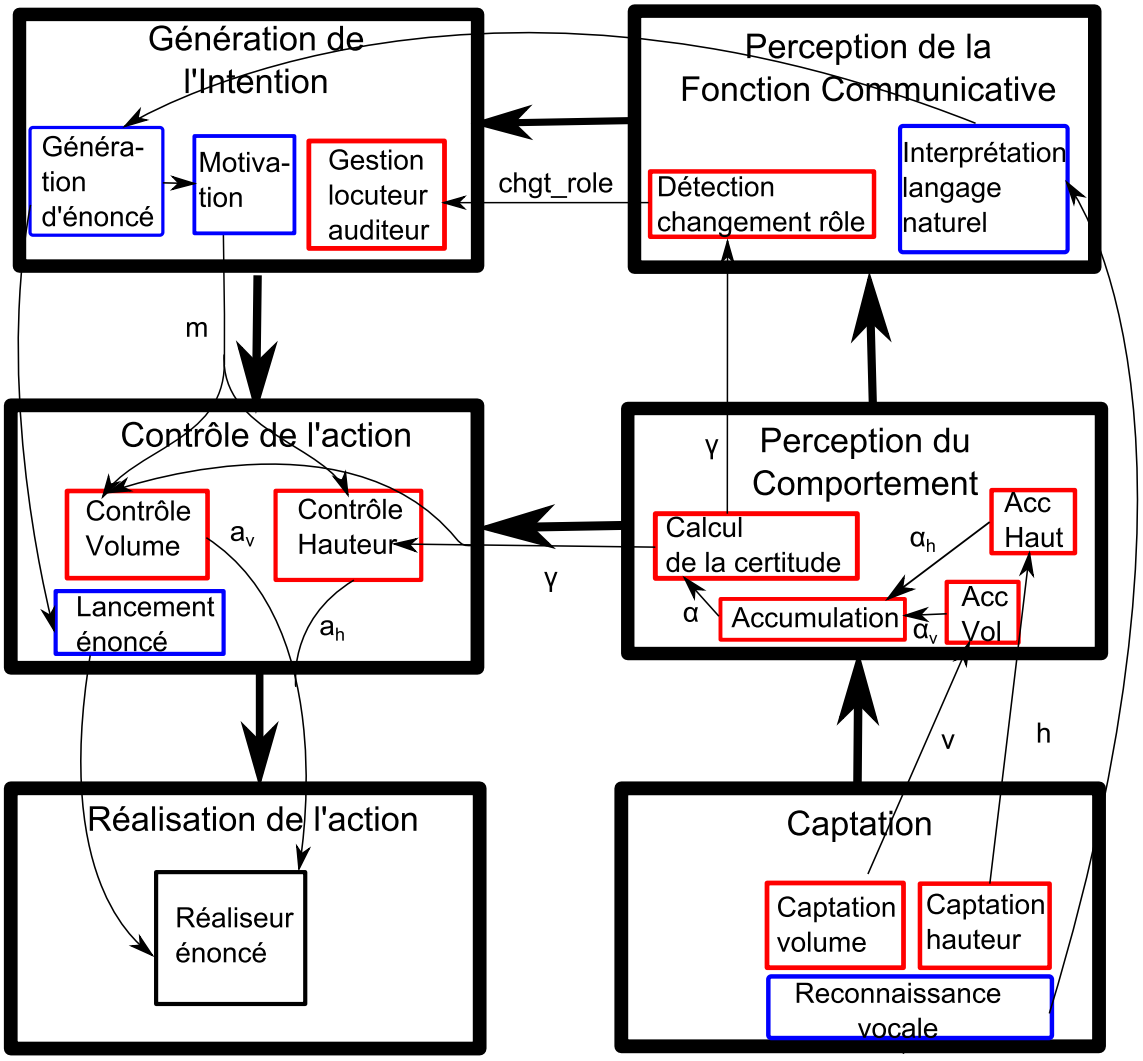


FIGURE 10.3 – Illustration de l'architecture en tenant compte de la relation entre les modules impliqués dans la gestion du tour de parole et les modules impliqués dans la gestion du dialogue.

perception continue du volume sonore et de la hauteur de voix a été réalisée en utilisant l'outil OpenSmile (Eyben *et al.*, 2013). L'outil fonctionne dans un processus parallèle au processus exécutant l'architecture d'agent et envoie les données en TCP à l'architecture. La fréquence de mesure et d'envoi des données de volume sonore et de hauteur de voix a été fixée à 100 Hz. L'agent dispose enfin d'une représentation graphique. La représentation graphique du personnage (personnage réalisé par la société Rocketbox) est implémentée dans le moteur de jeu Unity3D. Concernant les signaux non-vocaux liés à cette représentation graphique, le personnage dispose pour l'instant simplement d'une animation de parole et d'une animation d'écoute. Le lancement de ces animations est géré par un réaliseur implémenté dans le moteur de jeu en C#.

L'exploitation des capacités de notre modèle à supporter les recouvrements compétitifs nécessite de disposer d'une reconnaissance vocale et d'une synthèse vocale incrémentale. Avec une reconnaissance vocale incrémentale et l'utilisation d'un ensemble fini d'énoncés que l'agent est capable de reconnaitre, l'agent peut formuler très tôt une hypothèse sur l'énoncé que l'utilisateur est en train de produire. L'agent n'est donc pas obligé d'attendre la fin de la production de l'énoncé par l'utilisateur pour le traiter et programmer la génération de son énoncé. De ce fait, il peut potentiellement interrompre l'utilisateur. Nous avons utilisé le module de reconnaissance vocal de Microsoft. Pour la synthèse vocale incrémentale, l'outil inpro_iSS élaboré par Baumann (2013) a été utilisé.

Nous illustrons sur la figure 10.4 la succession des étapes menant à la génération et au lancement d'un énoncé dans le synthétiseur vocal.

L'agent est actuellement capable de synchroniser l'animation de parole dans Unity3D et le lancement de l'énoncé dans la synthèse vocale inpro_iSS. Cette synchronisation est réalisée par un décideur du module de contrôle de l'action qui, à chaque exécution, récupère l'état (START, INTERRUPTED ou END) de l'exécution de l'énoncé dans le synthétiseur vocal et envoie en retour une commande d'animation de parole ou de fin d'animation de parole selon l'état de l'action au gestionnaire d'action. Le réaliseur lié au synthétiseur vocal de l'agent peut recevoir deux commandes motrices différentes, une commande motrice « say » possédant comme paramètre la phrase à générer par l'agent et une commande motrice « changeProsody » possédant deux paramètres de contrôle pour le volume sonore et la hauteur de voix. Les équations de contrôle de l'action ont été reprises des équations établies au chapitre 9. Les équations de perception du comportement de l'agent ont néanmoins été modifiées pour tenir compte des défauts observés dans la récupération des signaux prosodiques par openSmile. Nous avons ainsi accentué les paramètres de détection de la présence de volume sonore et de hauteur de voix pour compenser la détection erronée de moments où le participant parle comme des moments de silence.

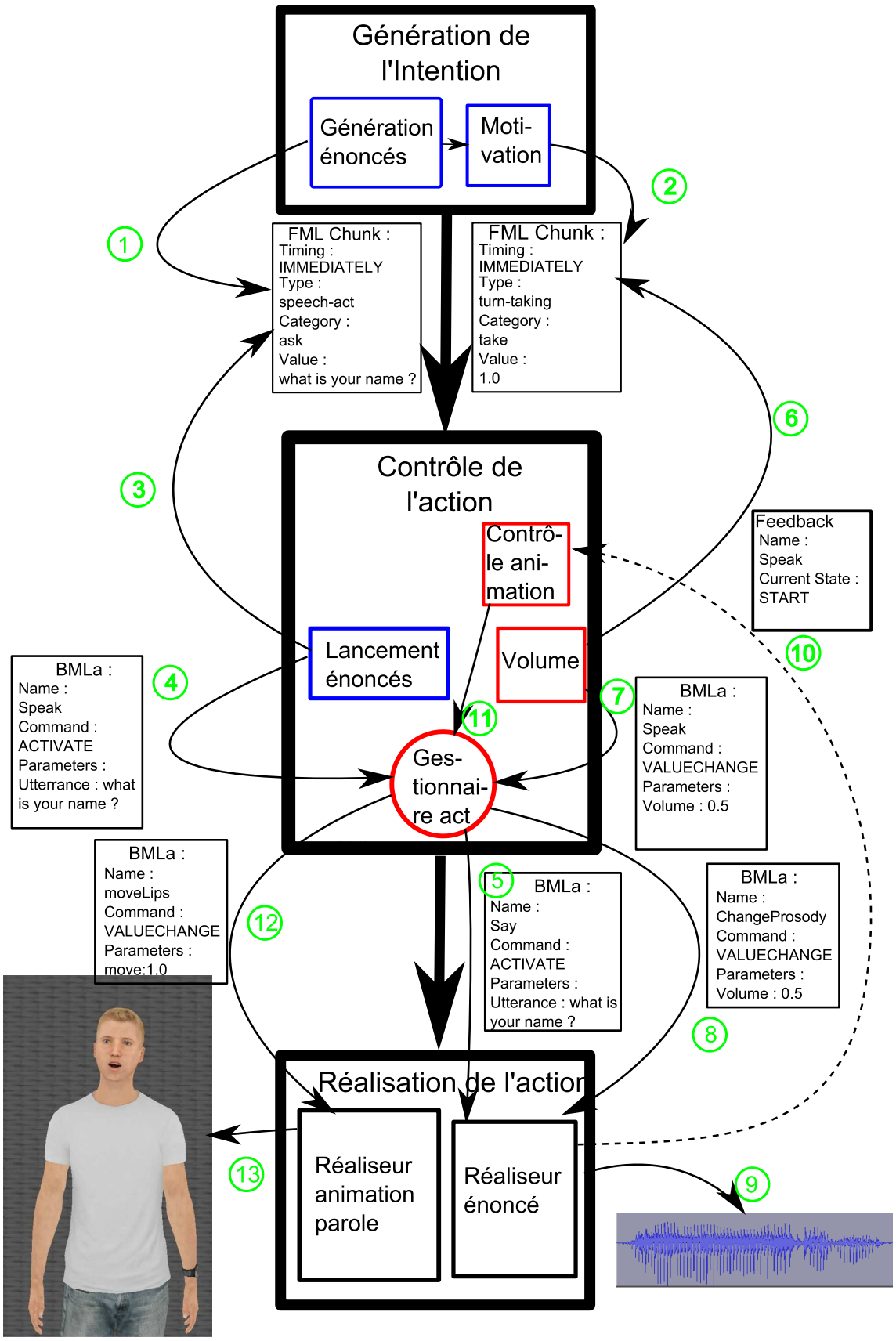


FIGURE 10.4 – Illustration de la communication entre les modules dans le cas d'une prise de tour de l'agent.

La figure 10.4 illustre le fonctionnement de l'architecture dans un exemple où l'agent n'a pas la parole et prend le tour. Lorsque l'agent a une nouvelle phrase à prononcer, il produit une intention liée à cette phrase (étape 1). Cette intention est formulée sous forme de *chunk* FML renseignant la catégorie de l'acte communicatif et le type d'attribut en accord avec Cafaro *et al.* (2014). L'énoncé prononcé par l'agent est renseigné sous forme de paramètre de l'acte communicatif correspondant représenté par le champ *Value* sous la figure. L'agent cherche ici à demander le nom de l'utilisateur : il génère un acte de langage (*speech-act*) demandant une information à l'utilisateur (type *ask*). Le sous-module de génération d'énoncés informe le sous-module traitant la motivation à parler de la génération d'un énoncé. Le sous-module de contrôle de la motivation à parler génère alors un *chunk* FML renseignant une valeur de motivation à 0.1 indiquant que l'agent veut parler (étape 2). Les deux *chunk* sont stockés dans le *blackboard* du module de contrôle de l'action. Le *chunk* FML spécifiant l'acte de langage est interprété par le sous-module du module de contrôle de l'action chargé de transformer la motivation de l'agent en requête d'action envoyée au gestionnaire d'action (étape 3). La contrainte de temps étant définie à la valeur *IMMEDIATELY*, le sous-module chargé du lancement d'action transforme directement la motivation de l'agent en commande d'action (étape 4). Cette commande d'action est alors envoyée sous forme de commande *BML* au gestionnaire d'action, qui se charge de faire correspondre la requête d'action à une commande motrice, ici la commande *say* et envoie cette commande au réaliseur d'énoncé (étape 5). La manière dont le réaliseur d'énoncé exécute cette commande dépend de la valeur de volume sonore actuelle de l'agent. Si le volume sonore est en dessous du seuil d'activation, le réaliseur mémorise l'énoncé mais ne l'envoie pas au synthétiseur vocal, à l'inverse si la valeur de volume sonore est supérieure au seuil, le réaliseur envoie l'énoncé au synthétiseur vocal. Dans ce cas de figure, le synthétiseur interrompt la phrase (en attendant la fin du mot qui est en train d'être prononcé) qu'il est en train de prononcer si c'est le cas, avant de lancer la phrase envoyée par le réaliseur. Dans notre exemple, la valeur du volume sonore lorsque le réaliseur reçoit la commande motrice est à 0 : celui-ci mémorise la phrase à prononcer mais ne l'envoie pas au synthétiseur vocal. Pendant ce temps, le décideur chargé de contrôler le volume sonore (module *Volume* en rouge sur la figure) module le volume sonore selon la nouvelle valeur de motivation générée à l'étape 2. Il exécute l'équation de contrôle du volume sonore puis envoie la valeur de sortie au gestionnaire d'action à chaque pas de temps. Imaginons maintenant que l'utilisateur vient de s'arrêter de parler. L'agent ayant une valeur de motivation favorisant la prise de tour et la certitude de l'agent convergeant vers l'alternative en faveur d'un abandon de tour de l'utilisateur, le volume sonore augmente. Le volume sonore et la hauteur de voix de l'utilisateur étant toujours à 0, la valeur de volume sonore finit par franchir le seuil d'activation. Comme à chaque pas de temps, une commande d'action

renseignant la valeur du volume sonore est transmise au gestionnaire d'action (étape 7) qui transforme la commande d'action en commande motrice à l'aide du lexique moteur puis transmet la valeur mise à jour du volume sonore au réaliseur d'action (étape 8). Le volume sonore ayant dépassé le seuil d'activation, le réaliseur d'action transmet l'énoncé au synthétiseur vocal, qui commence à générer et jouer le flux audio (étape 9). L'état de l'action du réaliseur passe alors à START et le réaliseur envoie un *feedback* renseignant la nouvelle valeur de l'état de l'action au gestionnaire d'action. Le sous-module chargé de contrôler le lancement de l'animation de parole (*Contrôle animation* sur la figure), récupérant en entrée l'état de la commande de parole (étape 10) génère alors une commande d'exécution de l'animation de parole (étape 11) qui est transformée en commande motrice par le gestionnaire d'action (étape 12) et envoyée au réaliseur chargé de contrôler les animations de parole et d'écoute de l'agent. Le réaliseur lance alors l'animation de parole (étape 13).

10.2 Validation du modèle

10.2.1 Respect des scénarios de dialogue

L'intérêt de notre modèle est d'assurer la coordination des échanges de parole au-delà du strict tour par tour (un locuteur à la fois, pas de silence, pas de recouvrement). Le tableau 10.1 présente un ensemble de scénarios d'interaction qu'un bon gestionnaire de tours de parole doit être capable de supporter dans le contexte d'un dialogue à initiative mixte. Pour réaliser cette étude, nous avons implémenté un gestionnaire de dialogue extrêmement simple qui repose sur le principe des paires adjacentes. Quand l'agent ne comprend pas ce que dit l'utilisateur, s'il est interrompu, alors il continue la production de son énoncé, et s'il prend le tour à la suite de l'utilisateur alors il produit un énoncé quelconque. Nous illustrons le résultat de l'exécution des différents scénarios avec notre agent sur les figures 10.5, 10.6,10.7 et 10.8. Sur chaque figure, l'évolution temporelle du volume sonore de l'agent est représentée par la courbe en vert. Nous avons également fait figurer les transcriptions des énoncés échangés entre l'agent (haut) et l'utilisateur (bas). Notons que notre agent s'exprime en anglais car la voix française produite par le synthétiseur vocal incrémental n'était pas d'une qualité suffisante. Pour les scénarios 2.1 et 2.2, l'annotation « detect » indique le moment où l'agent a interprété l'énoncé de l'utilisateur et planifie sa réponse.

La figure 10.5 montre l'exécution du scénario 1.1. L'agent prononce un énoncé et l'utilisateur tente de l'interrompre. L'agent veut absolument garder la parole (m=-1.0). Dans ce cas, l'agent repère rapidement la tentative d'interruption de l'utilisateur et commence à augmenter le volume de sa voix sur les expressions « mount » et « a shelf » et sa voix reste forte jusqu'à « red ». Ceci s'explique par une variable de certitude devenant positive lorsque l'utilisateur parle pour le locuteur courant.

Interruption de l'utilisateur	Scénario 1.1	L'utilisateur cherche à interrompre l'agent, mais l'agent prononce un énoncé qu'il considère comme très important et veut que l'utilisateur l'écoute jusqu'au bout : il augmente son niveau sonore
	Scénario 1.2	L'utilisateur cherche à interrompre l'agent, l'agent prononce un énoncé qu'il considère comme peu important, avec une motivation faible et laisse la parole à l'utilisateur en s'interrompant
Interruption de l'agent	Scénario 2.1	L'agent a compris ce que voulait dire l'utilisateur, a une volonté forte de parler et n'attend pas que l'utilisateur ait fini de parler pour l'interrompre
	Scénario 2.2	L'agent a compris ce que voulait dire l'utilisateur, il ne prend la parole que lorsque l'utilisateur a fini de parler

TABLE 10.1 – Quatre scénarios d'interaction couverts par notre modèle.

Cette variable de certitude positive est transmise du module de perception du comportement au module de contrôle des actions. Les sous-modules chargés de contrôler les actions de l'agent augmentent alors leur attracteur du fait de l'augmentation de la valeur de confiance. S'ensuit alors une augmentation des attributs prosodiques, qui est transmise au réaliseur chargé de contrôler l'exécution de l'énoncé actuel de l'agent par l'outil de synthèse vocale. Ce réaliseur envoie alors la modification du volume sonore et de la hauteur de voix à la synthèse vocale, résultant en une voix plus forte et plus aigüe.

La figure 10.6 montre l'exécution du scénario 1.2. L'agent prononce un énoncé, et l'utilisateur tente de l'interrompre. En comparaison au scénario 1.1, ici l'agent a une motivation faible à parler ($m = -0.1$), ce qui implique, au regard des équations contrôlant l'évolution des attracteurs des attributs prosodiques, une baisse du volume sonore et de la hauteur de voix à leur valeur minimale 0. La valeur de volume sonore étant inférieur au seuil de silence fixé le flux de synthèse vocale est ici interrompue à la fin du mot actuellement prononcé par l'agent. L'agent s'arrête alors de

parler en réponse à la tentative d'interruption. Lorsque l'utilisateur s'arrête de parler après l'interruption, s'ensuit un moment de silence. Pendant ce moment de silence, la variable de certitude augmente de nouveau, résultant alors en une augmentation des attracteurs des équations de contrôle des attributs prosodiques. Les valeurs de volume sonore et de hauteur de voix se mettent à nouveau à augmenter. Lorsque le volume sonore dépasse la valeur seuil de silence du réaliseur, ce dernier charge le synthétiseur vocal de reprendre l'exécution de l'énoncé là où il s'était arrêté.

La figure 10.7 montre une interaction entre l'utilisateur et l'agent correspondant au scénario 2.1. L'utilisateur pose une question à l'agent, et l'agent reconnait la phrase prononcée par l'utilisateur avant la fin. Une fois que l'agent a reconnu ce que dit l'utilisateur, il l'interrompt. Cela est possible par le fait que dans tous les cas de figure, lorsque l'outil de reconnaissance vocale reconnait la phrase prononcé par l'agent avant la fin avec un niveau de confiance élevé, le module générant les énoncés de l'agent sélectionne la réponse. Cette réponse est transmise au module de contrôle de l'action, puis traduite en requête d'action et transmise par le gestionnaire d'action au réaliseur contrôlant le synthétiseur vocal. Le lancement ou non de l'énoncé dans le synthétiseur vocal dépend alors des valeurs de volume sonore et de hauteur de voix. Ces valeurs sont ici élevées, provoquant un déclenchement de l'énoncé avant la fin de la phrase de l'utilisateur. L'agent a dans le cadre de ce scénario une motivation de $m = 1.0$ indiquant la plus forte motivation à prendre le tour.

La figure 10.8 montre la réalisation du scénario 2.2. L'utilisateur fournit la même entrée que le scénario 2.1, mais l'agent a dans ce scénario une motivation de prendre le tour de $m = 0.1$. Par rapport au scénario 2.1, lorsque la phrase est transmise au réaliseur, les valeurs de volume sonore et de hauteur de voix font que la phrase n'est pas lancée dans le synthétiseur vocal. L'agent attend alors d'être certain ($\gamma = 0.9$) que l'utilisateur a laissé le tour avant de prendre la parole ce qui survient plus d'une seconde après la fin de tour de l'utilisateur.

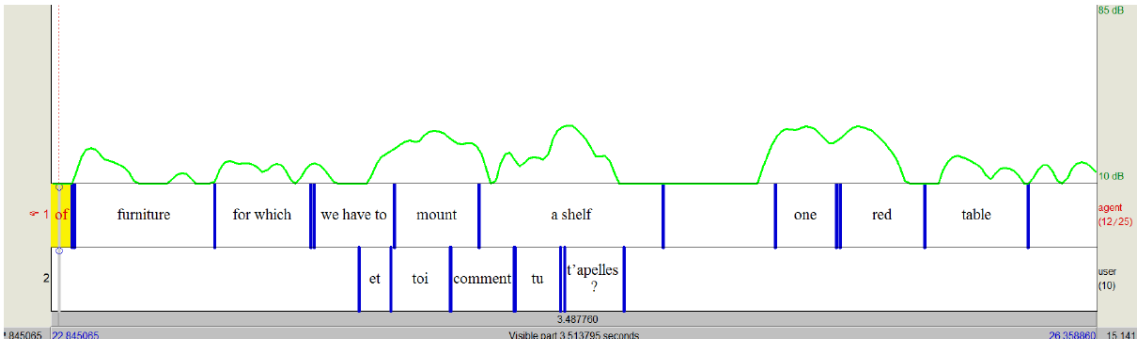


FIGURE 10.5 – Scénario 1.1.

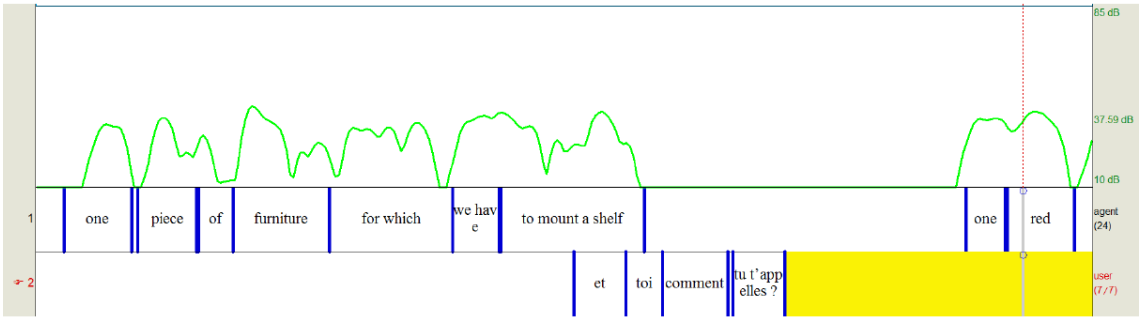


FIGURE 10.6 – Scénario 1.2.

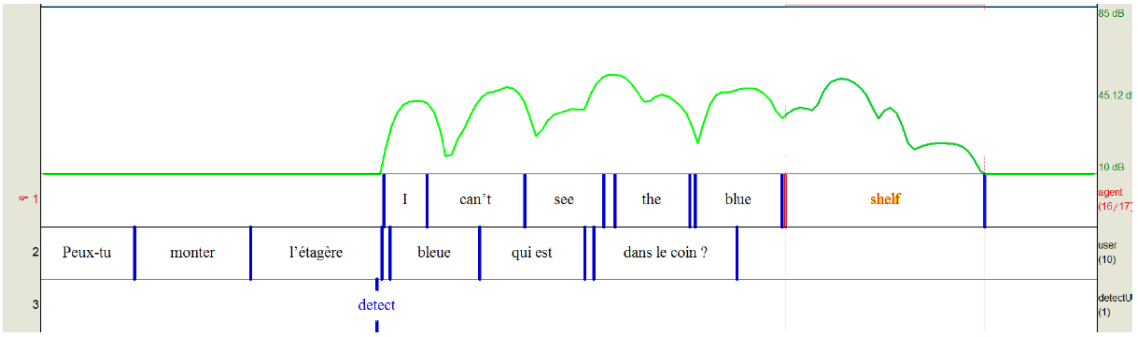


FIGURE 10.7 – Scénario 2.1.

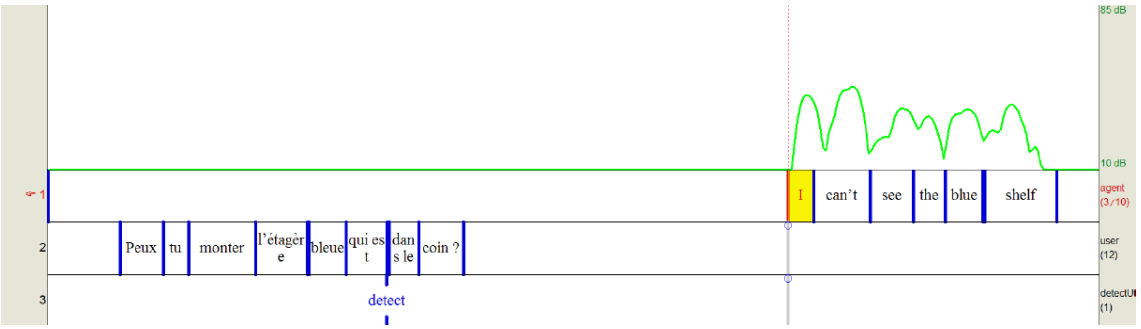


FIGURE 10.8 – Scénario 2.2.

10.2.2 Évaluation du modèle dans le cadre d'une interaction utilisateur-agent

Présentation du protocole

Nous proposons de comparer notre modèle de tour de parole avec une implémentation d'un second modèle. Dans ce second modèle, la prise de parole de l'agent est dirigée par des règles simples où l'agent ne prend la décision de parler qu'après un intervalle de temps suivant la fin de tour de l'utilisateur, et s'interrompt systématiquement lorsque l'utilisateur parle en même temps que l'agent, approche très souvent employée dans les architectures d'agent et considérée comme non-optimale (Ward *et al.*, 2005). Nous reprenons les valeurs seuils utilisées dans la littérature, ainsi l'agent attend dans notre cas 600 ms après la fin de tour de l'utilisateur pour parler, et ne détecte un tour de parole de l'utilisateur qu'à partir de 100 ms après

avoir détecté la voix de l'utilisateur. Nous comparons cette implémentation avec notre modèle d'échange de parole où l'agent varie sa motivation de parler au cours de la conversation avec l'utilisateur. Pour systématiser les variations de stratégie de prise de parole au cours de l'interaction, nous avons décidé de diviser celle-ci en deux parties : dans la première partie, l'agent a une motivation faible de parler, s'interrompant lorsque l'utilisateur lui coupe la parole et attendant que l'utilisateur ait fini de parler pour commencer à parler. Dans la seconde partie, l'agent a une motivation forte de parler impliquant que dès que ce dernier a quelque chose à dire il cherche à interrompre l'utilisateur et ne laisse jamais l'utilisateur prendre la parole. À des fins de simplification, dans la suite de l'article, nous nommerons la condition ou les prises de parole sont pilotées par notre modèle « condition 1 », comprenant elle-même deux parties, la « condition 1 Weak », condition ou l'agent a une motivation faible de parler et la « condition 1 Strong », condition ou l'agent a une motivation forte. La condition correspondant au contrôle du tour de parole a base de temporisations est notée ici « condition 2 ». Pour l'interaction dialogique utilisateur-agent, nous avons employé le même scénario de négociation que nous avions utilisé pour l'analyse des interactions humaines. Le participant pense qu'il est proche de la côte, l'agent pense, lui, qu'il est loin de la côte.

Afin de ne pas limiter les phrases que peut dire l'utilisateur, nous avons remplacé le composant de reconnaissance vocale par un magicien d'oz. Ce dernier interprète la phrase de l'utilisateur et décide de la phrase la plus appropriée à générer ensuite. Pour évaluer les différences d'efficacité des deux modèles, il est nécessaire pour le magicien d'oz de choisir la phrase à générer avant que l'utilisateur ait fini de parler, afin de laisser aux modèles de tour de parole le contrôle du moment où l'agent commence à parler. Il faut pour cela une interface de contrôle limitant au maximum le temps nécessaire pour sélectionner une phrase. L'interface de choix des phrases est montrée sur la figure 10.9.

L'interface est décomposée en quatre parties, la partie 1 contient les boutons générant les phrases à prononcer par l'agent. Les commandes sont décomposées en arbre, en haut de l'arbre se trouve les croyances de l'agent sur sa situation actuelle (« Loin de la côte » ou « près de la côte »). Étant donné les croyances de l'agent, celui-ci dispose de plusieurs désirs liés à sa croyance, ces désirs composent la deuxième couche de boutons. Ainsi s'il pense être loin de la côte, il a comme désir de pouvoir se nourrir pendant plusieurs semaines et de pouvoir se réchauffer. S'il pense être près de la côte, il va donner la priorité à pouvoir être repéré et à rejoindre la côte. Selon ses désirs, il voudra prendre des objets correspondant à ses désirs (« prendre un kit de pêche » ou « prendre des rations » s'il veut pouvoir se nourrir, et « prendre des couvertures » s'il veut pouvoir se réchauffer). À chaque bouton est associé un ensemble de phrases de sorte que deux clics consécutifs sur le même bouton ne génèrent pas la même phrase. L'ensemble des phrases générées

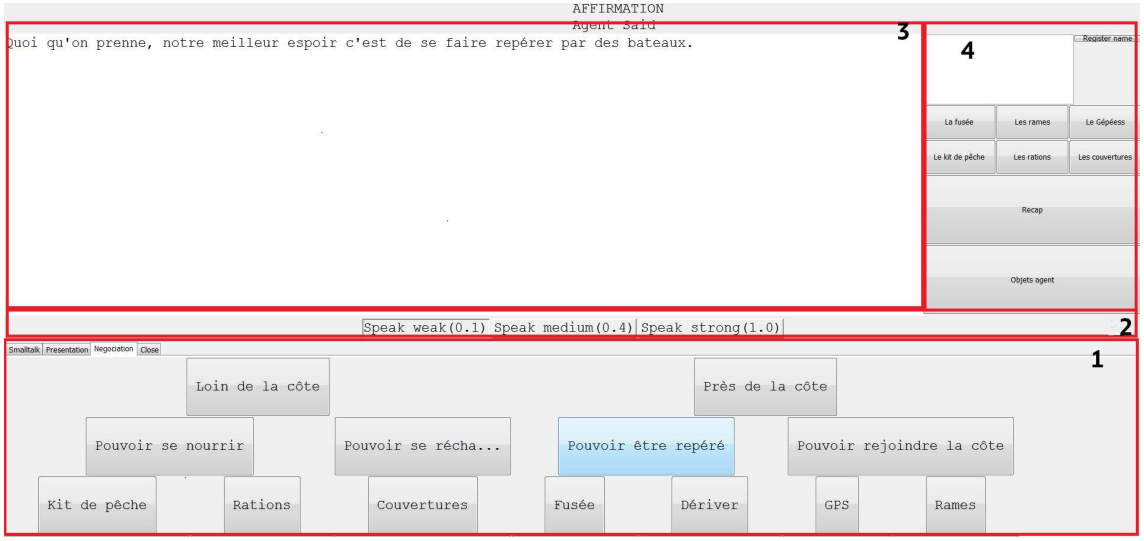


FIGURE 10.9 – Interface utilisée par le magicien d'oz pour générer les énoncés de l'agent.

par l'interface proviennent de notre corpus d'interactions humaines. Pour chaque bouton, le magicien d'oz dispose d'arguments et de contre-arguments. Par appui sur une touche du clavier il peut facilement basculer d'une interface lui permettant de générer des arguments en faveur des objets à une interface générant des arguments en défaveur des objets. Si l'on considère par exemple l'objet « fusée », l'un des arguments proposés est : « La fusée de détresse je pense que c'est un bon moyen de se faire repérer ». L'objet fusée dispose aussi de plusieurs contre-arguments dont : « une fusée de détresse, le problème c'est que si elle n'est pas vue on ne peut plus en relancer une après ». La structure en arbre a pour objectif de simplifier la recherche de la phrase à prononcer du point de vue du magicien d'oz en fournissant une structure logique à l'ensemble des phrases à générer. En plus des phrases présentes ici le magicien d'oz a possibilité de générer des *backchannels* (« d'accord », « ok »), des réponses à des questions fermées (« oui », « non ») et signifiant son accord ou son désaccord (« effectivement », « je suis tout à fait d'accord », « je ne suis pas du tout d'accord », par exemple), et de demande de répétition (« je n'ai pas compris peux-tu répéter ? »). Afin de gagner du temps sur la génération de phrase, le magicien d'oz génère une phrase en cliquant simplement sur le bouton correspondant. Le magicien d'oz n'a ainsi besoin que d'une action pour générer une phrase. Pour permettre de vérifier que ce dernier a bien sélectionné la bonne phrase, dès qu'il survole le bouton, le magicien d'oz peut voir ce que l'agent va générer comme phrase, dans la zone de texte dans la partie 3 de l'interface. La partie 2 contrôle les motivations de l'agent, il aura ainsi une faible ou forte motivation de parler, résultant en un agent favorisant la prise de parole de l'utilisateur ou interrompant l'utilisateur pour prononcer sa phrase. La partie 4 permet enfin de fournir à l'utilisateur un retour sur l'ensemble des objets sur lesquels les participants se sont mis d'accord, et sur les objets que l'agent compte prendre. Chaque participant interagit deux fois avec l'agent, chaque

Question	Médiane condition 1	Médiane condition 2	p-value
"Ne percevait pas les moments où je parlais"	2.25	1.75	0.95
"Prenait la parole aléatoirement"	2.5	2.625	0.6
"M'a coupé la parole involontairement"	6	4	0.019*
"M'a parfois délibérément coupé la parole"	6	6.5	0.91
"A fait attention à ne pas me couper la parole"	4.5	7.5	0.006**
"A mis du temps à me répondre"	3	2	0.77
"A parfois refusé de me laisser parler"	6.125	5.75	0.16
"J'ai été gêné par les prises de paroles de mon interlocuteur"	4.5	3.25	0.54
"Je me suis senti à l'aise dans le dialogue"	5.25	6.25	0.55
"J'ai aimé parler avec mon interlocuteur"	6.625	7	0.52
"Le comportement de mon interlocuteur était proche d'un comportement humain"	5.625	6.5	0.97

TABLE 10.2 – Questions et réponses des participants pour les conditions 1 et 2.

fois avec une condition différente et selon le même scénario. Chaque interaction dure deux minutes trente secondes. À la fin de chaque interaction, un questionnaire est proposé au participant évaluant entre autres sa satisfaction à interagir avec l'agent, sa facilité d'interaction et sa perception du caractère intentionnel ou non des interruptions par l'agent. Dans ce questionnaire, différentes affirmations sont présentées au participant, le participant renseigne son niveau d'accord avec l'affirmation entre pas du tout d'accord et tout à fait d'accord sur une échelle continue de 0 à 10. Les affirmations ont été inspirées de Skantze et Hjalmarsson (2010), Bevacqua *et al.* (2014) et De Vault *et al.* (2015).

Résultats

31 étudiants, ingénieurs et chercheurs (30 hommes, une femme) ont en tout participé à l'expérimentation. Tous avaient pour langue maternelle le français.

Les résultats au questionnaire sont montrés sur le tableau 10.2. Peu de différences

significatives sont observées entre les conditions. Les participants aiment globalement parler avec l'agent (médiane de 7). Les résultats en termes de crédibilité sont plus mitigés, la médiane à l'affirmation "Le comportement de mon interlocuteur était proche d'un comportement humain" étant de 6 seulement. Les utilisateurs perçoivent bien que l'agent a fait attention à ne pas leur couper la parole au cours de l'interaction ($p=0.006$) dans la condition 2 (médiane de 7.5) par rapport à la condition 1 (médiane de 4.5). Néanmoins, les participants perçoivent contre-intuitivement que l'agent leur a plus coupé involontairement la parole ($p=0.018$) dans la condition 1 (médiane de 6) par rapport à la condition 2 (médiane de 4).

L'attribution par l'utilisateur d'un caractère involontaire aux coupures de parole dans la condition 1 pose problème. Notons que le nombre de recouvrements par l'agent observés est bien supérieur à ce qui était attendu : en effet, ce nombre n'est pas significativement différent selon les conditions 1 et 2. Nous pensons que cela pourrait s'expliquer par une perception des interruptions volontaires de l'agent dans la condition 1 « Strong » comme des erreurs de coordination de l'agent.

L'évaluation subjective de l'interaction par un utilisateur pouvant être sensible à des biais liés notamment à la formulation des questions, des analyses objectives des interactions ont été réalisées en complément. Une analyse des durées de transition entre les différentes conditions (condition 1 Weak, condition 1 Strong, condition 2) à la fois utilisateur-agent et agent-utilisateur, a été réalisée en complément. La répartition des durées de transitions utilisateur-agent est montrée sur la figure 10.10, et la répartition des durées de transitions agent-utilisateur sur la figure 10.11.
Les résultats montrent des transitions moyennes agent-utilisateur plus courtes pour la condition 1 par rapport à la condition 2 (1.39 s en moyenne pour la condition 2 et 1.11 s pour la condition 1). Néanmoins aucune différence significative n'a été observée entre la condition 1 « Weak » et la condition 1 « Strong ». Pour les transitions utilisateur-agent on obtient de même une différence significative ($p<0.05$) entre la condition 1 (0.84 s) et la condition 2 (1.19 s).
En complément des valeurs de durée de transition agent-utilisateur, la variation de hauteur de voix (hauteur de voix du participant) de chaque participant a été mesurée lors des moments de recouvrement. Les résultats montrent une variation de hauteur de voix ($p=0.034$) significativement supérieure lors des moments de conflits par rapport à la valeur moyenne de hauteur de voix du participant pour la condition 1 Strong, mais ne montre pas de différence dans la valeur de hauteur de voix lors des moments de conflits entre les trois conditions.

10.2.3 Discussion

Les résultats de notre questionnaire ne semblent pas montrer d'effets de la variation des stratégies de prise de parole sur le ressenti de l'utilisateur au sujet de l'interaction. Ces données de questionnaires ont été croisées avec des analyses com-

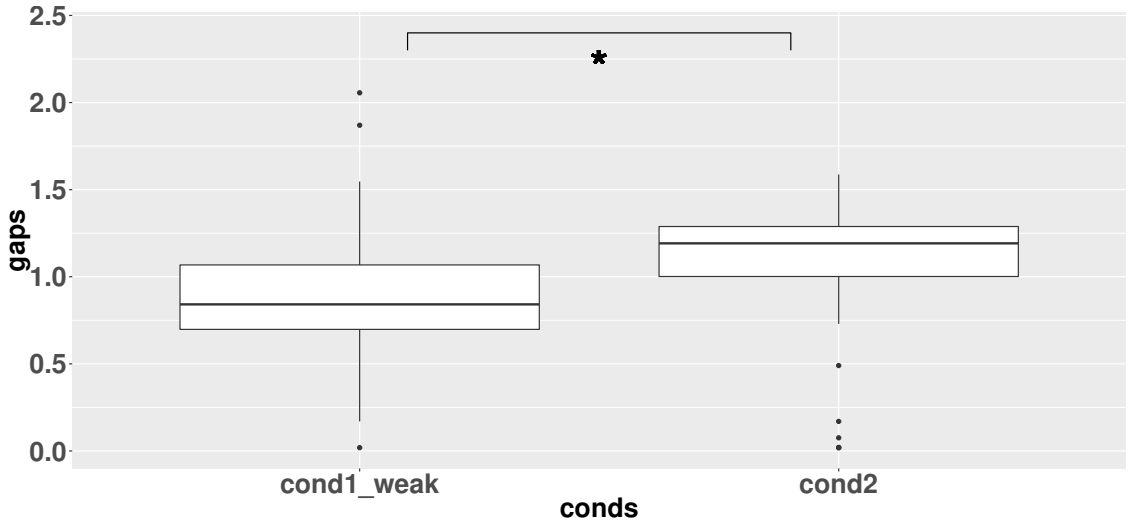


FIGURE 10.10 – Répartition des durées de transitions utilisateur-agent.

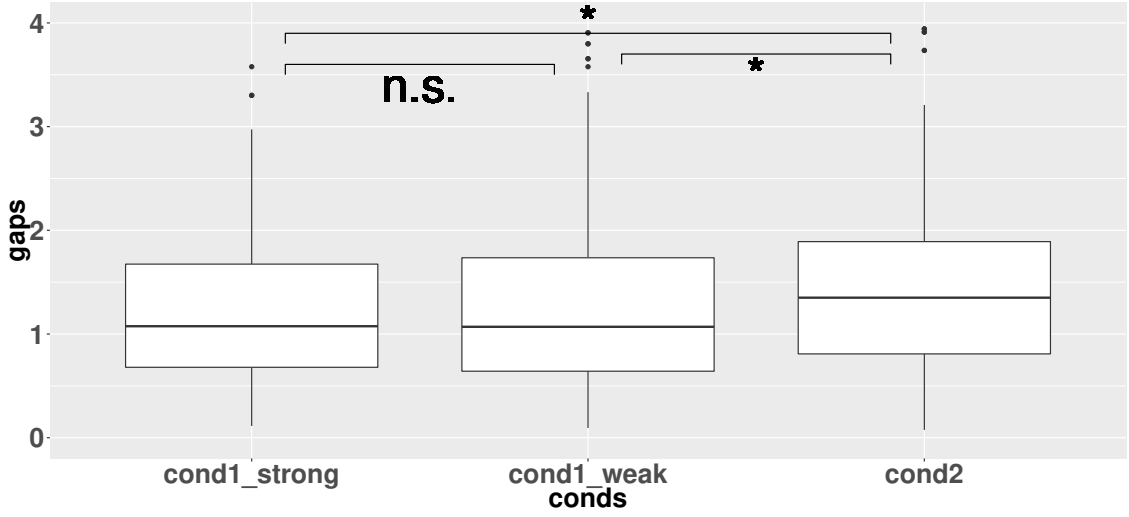


FIGURE 10.11 – Répartition des durées de transition agent-utilisateur.

portementales montrant une réaction de l'utilisateur en termes de hausse de hauteur de voix lors de la coupure de parole de l'agent, mais pas de modification du temps de prise de parole de l'utilisateur. Nous avons recueilli, à la fin de l'expérimentation, les impressions orales des participants sur l'interaction. Ceux-ci ont rapporté des impressions moins catégoriques face aux coupures de parole que ce qui a été observé lors de l'analyse du questionnaire. Six participants ont ainsi explicitement mentionné qu'ils avaient perçu les coupures de parole comme involontaires alors que treize participants ont perçu au moins certaines coupures comme volontaires. Parmi ces treize participants, quatre participants ont jugé ces coupures comme justifiées, pertinentes ou normales, et cinq participants ont associé ces coupures au fait que l'agent n'était pas d'accord ou cherchait à imposer ses idées. Enfin, cinq de ces participants ont associé un caractère « humain » à ces coupures. Néanmoins, cette perception des coupures n'amène pas à un meilleur ressenti de l'interaction de la part de l'utilisateur. Un des participants a rapporté un sentiment de « rage » de s'être fait couper la parole plusieurs fois, et deux autres participants ont rapporté que ces coupures représentaient une gêne dans l'interaction. Les résultats du questionnaire ont aussi montré que les sujets percevaient l'agent comme réactif et ont peu noté la présence de silences « génants » dans la conversation. Lorsque ces moments de silences étaient perçus, ils étaient considérés comme peu naturels, bien que deux participants ont mentionné ces moments comme crédibles et liés à un agent qui réfléchissait à ce qu'il voulait dire. Ce caractère non-naturel est peut-être lié au fait que l'interaction était uniquement audio avec l'agent empêchant à l'agent de fournir une rétroaction visuelle à l'utilisateur.

L'erreur de détection élevée (50 % des transitions utilisateur-agent) peut être en partie expliquée par la détection de la voix de l'utilisateur, montrant un nombre important de moments où aucune voix n'est détectée alors que l'utilisateur parle. Le fait que les participants sont peu dérangés par les temps de silence moyens importants observés dans le cadre de l'interaction étaye le fait que ces derniers n'attendent pas de prises de parole optimales de la part de l'agent dans le cadre de scénarios d'initiative mixte. Le caractère approprié ou non des coupures de parole semble plus sujet à discussion. Les témoignages recueillis des participants tendent à montrer que ces coupures peuvent donner une impression de fluidité et d'immersion dans le dialogue à condition que la coupure soit appropriée au contexte du dialogue. Enfin le manque de distinction entre des coupures de parole délibérées et non délibérées peut être liée à la qualité de la voix, jugée « mauvaise » par la majorité des participants, rendant peu naturelle l'augmentation du volume sonore de l'utilisateur.

10.3 Conclusion

Nous avons présenté dans ce chapitre une implémentation des équations de notre modèle dans l'architecture BeAware présentée chapitre 8. Nous avons démontré en premier lieu le fonctionnement du modèle dans un scénario de dialogue en temps réel avec l'utilisateur, en vérifiant la capacité de l'agent à respecter quatre scénarios de dialogue mentionnés dans la section 10.2.1. Nous avons ensuite présenté un protocole expérimental avec pour objectif de valider la contribution du modèle à l'amélioration de l'expérience de l'utilisateur en interaction avec le modèle. Aucune différence dans les réponses des participants aux questions sur la satisfaction suscitée par l'interaction, la crédibilité de l'agent et le sentiment de présence sociale générée par l'agent n'a été remarquée. Néanmoins, l'observation de données objectives sur le comportement de l'utilisateur envers l'agent montre un effet de la condition sur les prises de parole de l'utilisateur. Ce résultat nous laisse avec un certain nombre d'hypothèses à explorer sur la cause exacte de la variation du comportement de l'utilisateur selon la condition. Les retours oraux des participants sur leur impression montrent, eux, la capacité des participants à s'immerger dans l'interaction. Ils n'éprouvent ainsi aucune difficulté à attribuer des comportements anthropomorphiques à l'agent et réagissent de manière émotionnelle (sentiment de « rage ») au comportement de l'agent lorsque ce dernier coupe la parole. L'utilisateur a dans ce cas des réactions négatives vis-à-vis du comportement de l'agent, néanmoins, cela n'implique pas l'échec de l'interaction de l'utilisateur avec le système. Dans un scénario de formation en réalité virtuelle, l'objectif n'est pas nécessairement de susciter la satisfaction de l'utilisateur mais de reproduire une situation réelle. L'utilisateur pourrait ainsi être confronté dans une situation de travail collaboratif à des agents virtuels ayant différentes attitudes (amicales, hostiles) avec l'utilisateur. Dans ce cadre, l'utilisation de notre modèle de gestion du tour de parole avec un agent modulant la force de la motivation avec laquelle il souhaite prendre, laisser, garder la parole pourrait être intéressant, mais ici encore ce n'est qu'une hypothèse qui mérite d'être explorée.

L'interaction montre que la perception de la fin de tour n'est pas optimale dans notre architecture. Une première raison est liée à l'extraction des attributs prosodiques provenant d'openSmile, n'étant pas parfaits et éloignée des valeurs extraites « hors ligne » par des outils linguistiques comme Praat. Cependant, une partie de la cause de cette détection non optimale réside dans la manière dont notre modèle a été paramétré. Un apprentissage sur des extraits d'interactions humaines pourrait améliorer en ce sens la détection de la fin de tour de l'utilisateur. Cette non-optimalité reste néanmoins moins problématique dans une approche continue par rapport à une approche événementielle. Dans une architecture événementielle, une erreur dans la détection du comportement de l'utilisateur a un coût élevé puisque l'action de l'agent est plus difficile à interrompre une fois lancée. Au contraire, dans une ap-

proche continue, une prise de décision erronée peut être corrigée rapidement par la perception du comportement de l'utilisateur vis-à-vis de l'agent. Les erreurs de détection des fins de tour observées dans le protocole expérimental (50 %) doivent néanmoins être améliorées même pour un modèle continu. L'amélioration pourrait passer par un apprentissage des paramètres du module de perception du comportement en confrontant l'agent à des extraits d'interactions humaines. Cet apprentissage fournirait un jeu de paramètres de base à partir duquel l'agent s'adapterait à l'utilisateur.

Chapitre 11

Conclusion générale

Dans cette conclusion, nous résumons les différentes contributions réalisées au cours de cette thèse. Nous commençons par résumer notre contribution puis nous évaluons notre travail au regard des quatre questions générales posées en introduction de cette thèse. Nous concluons ce chapitre en présentant les améliorations possibles de notre modèle.

11.1 Synthèse des travaux

Au cours de cette thèse, nous avons élaboré un modèle pour la coordination des tours de parole dans une interaction dyadique utilisateur-agent. Nous avons d'abord conçu un modèle théorique s'appuyant sur deux modèles de psychologie cognitive : la dynamique comportementale de Warren (2006) et le modèle de dérive diffusion. L'originalité de ce modèle réside dans le caractère émergent de la coordination des tours de parole, créé par le couplage sensorimoteur continu entre les participants. Nous avons montré par une simulation entre deux agents les capacités d'adaptation de chaque participant à leur partenaire, et la robustesse des échanges de tour à une perception bruitée des signaux du partenaire. Nous avons ensuite implémenté notre modèle suivant une méthodologie centrée utilisateur : nous avons paramétré notre modèle à partir d'un corpus d'interactions humaines, puis nous avons évalué la capacité de notre modèle à coordonner ses échanges de tour avec l'utilisateur. Pour cette interaction avec l'utilisateur, nous avons conçu une architecture informatique supportant l'implémentation de processus de perception et d'action continus. Détaillons maintenant notre contribution au regard des questions générales énoncées en introduction.

Question générale 1 : Une approche émergente à base d'un modèle continu permet-elle de reproduire les comportements relatifs à la gestion des tours de parole dans des interactions humaines ? Pour répondre à cette question nous avons mené une expérimentation afin de valider certaines hypothèses

de conception du modèle et de mesurer les variations prosodiques, durées de transition et de conflit observées dans les interactions humaines. Les mesures effectuées nous ont permis de calibrer notre modèle à la fois manuellement et par un apprentissage artificiel. Les résultats permettent de s'assurer que nos agents sont capables de reproduire la plupart des comportements observés dans cette expérimentation. Ces résultats sont néanmoins à pondérer par plusieurs constats. Premièrement, les données sur lesquelles nous nous sommes appuyées pour paramétrer notre modèle proviennent d'un corpus de données d'une population particulière de participants, étudiants, doctorants, ingénieurs ou chercheurs d'une tranche d'âge définie qui devaient discuter selon un scénario de dialogue particulier. La question de savoir si les situations observées et les variations prosodiques et de durée de transition peuvent se généraliser à l'ensemble de la population est une question ouverte. Les résultats trouvés sont néanmoins en accord avec les résultats de Campione et Véronis (2002) pour les durées de transitions et en accord avec Gravano et Hirschberg (2011) pour les variations des signaux prosodiques, ces auteurs ayant réalisé leurs mesures sur des populations différentes laissant penser que ce résultat peut être généralisé.

Question générale 2 : À quel point notre agent est capable de s'adapter à son partenaire et à un environnement bruité ? L'analyse du modèle théorique a montré la capacité de deux agents contrôlés par notre modèle théorique à s'adapter à des partenaires ayant des temps de prises de décision et une inertie dans la production de leurs signaux variables. De même, une robustesse aux fluctuations aléatoires dans la perception et à la présence ou l'absence de signaux a été observée. Comme résultat principal de ces analyses d'interactions entre deux agents, il est important de noter que les capacités adaptatives des deux agents ne sont pas des principes explicitement renseignés à la création du modèle théorique, ni lors de la formulation des équations de perception et de contrôle des actions, ce sont au contraire des propriétés émergentes des interactions entre les agents. Ce résultat est prometteur pour l'application du modèle à des interactions utilisateur-agent et constitue une approche originale par rapport aux architectures de gestion du tour de parole habituelles, s'appuyant sur un apprentissage par renforcement pour améliorer leurs prises de parole. Néanmoins, l'application de ce modèle à une interaction temps-réel avec l'utilisateur a montré une difficulté de notre agent à s'adapter à l'utilisateur. Ces résultats montrent la limitation des capacités adaptatives de l'agent qui se dégradent rapidement, due en grande partie à l'imprécision des algorithmes de captation des signaux prosodiques utilisés. Nous présentons paragraphe 11.2.2 les solutions que nous avons envisagées pour améliorer l'adaptation de l'agent à l'utilisateur.

Question générale 3 : Un agent variant la manière dont il prend la parole modifie-t-il la perception que l'utilisateur a de l'agent ? Nous avons cherché à répondre à cette question en réalisant une expérimentation où l'utilisateur

était chargé d'interagir avec un agent prenant toujours le tour sans interrompre l'utilisateur et un agent pouvant interrompre ce dernier. Nous avons mesuré le ressenti de l'utilisateur entre les deux conditions et nous avons observé peu de variations du ressenti de l'utilisateur. Les résultats montraient, au contraire, que les participants ne percevaient pas les coupures de parole de l'agent comme délibérées mais plutôt comme une erreur de détection de la fin de tour. Lors d'entretiens avec l'utilisateur après l'expérimentation, nous avons néanmoins récolté des avis suggérant une perception des interruption de l'agent comme délibérés et liés à une intention communicative particulière de l'agent. La présence de ces avis en faveur d'une variation de la perception de l'agent nous pousse à réaliser un nouveau protocole expérimental afin d'explorer plus en détail les facteurs conduisant à une variation dans la perception des interruptions de l'agent.

Question générale 4 : À quel point notre modèle de gestion du tour de parole peut s'appliquer à un contexte réel de dialogue utilisateur-agent ? Pour l'adaptation du modèle à un contexte de dialogue réel nous avons créé une architecture d'agent inspirée de deux architectures existantes : ASAP (Kopp *et al.*, 2014) et Ymir (Thórisson, 1999) (voir chapitre 8). Le choix de s'appuyer sur des architectures existantes plutôt que de créer entièrement une architecture est motivé par la volonté de rendre intégrable notre modèle dans des architectures existantes d'agents disposant de capacités de gestion du contenu verbal plus avancé que ce que nous avons proposé, de gestion des émotions ou encore de réaliseurs gérant la gestuelle ou la direction du regard de l'agent. Le choix d'ASAP est en ce sens judicieux. En effet, en rendant compatible notre architecture avec le standard SAIBA nous facilitons à l'avenir l'intégration de travaux réalisés par d'autres laboratoires. Comme première étape, nous avons montré par l'implémentation du modèle dans l'architecture la capacité de notre module de gestion du tour de parole à fonctionner en complément d'une gestion du contenu.

11.2 Travaux futurs

11.2.1 Interprétation d'autres signaux

Pour l'instant, le modèle théorique été implémenté pour une coordination des tours de parole unimodale reposant uniquement sur les signaux prosodiques de volume sonore et de hauteur de voix. Avec cette implémentation, l'agent n'arrive pas à se coordonner en toutes circonstances avec l'utilisateur. Malgré les capacités adaptatives de notre modèle, permettant de compenser le manque de modalités par une accentuation des signaux que l'agent peut produire, permettre à l'agent d'interpréter et produire d'autres signaux non-verbaux améliorerait la coordination des tours de parole. Hjalmarsson (2011) a en effet montré que lorsque l'on demandait aux

participants dans une expérimentation de perception utilisateur de juger à partir d'un extrait de conversation si le locuteur courant allait laisser la parole ou non, plus le nombre d'indices non verbaux de fins de tour était grand, plus la probabilité que le participant juge cet extrait comme une fin de tour était grande. Bien que l'observation de Hjalmarsson (2011) n'ait pas été effectuée en condition écologique, cela laisse supposer que le même principe s'applique dans de réelles interactions dialogiques. Cette conclusion est renforcée par les résultats de Gravano et Hirschberg (2011), montrant une corrélation entre le nombre de signaux de fins de tour et le nombre de prises de tour du locuteur suivant. Parmi les signaux que l'agent pourrait interpréter figure en premier lieu le débit de parole du participant. Plusieurs auteurs ont en effet trouvé des corrélations entre les variations de débit de parole et les fins de tour des participants (Duncan, 1972; Gravano et Hirschberg, 2011). De même, les variations de direction du regard, très souvent utilisés dans le cadre du tour de parole utilisateur-agent (Lessmann *et al.* (2004) ou Skantze *et al.* (2014) présentent des exemples de systèmes utilisant le regard pour se coordonner), ou la gestuelle (Duncan, 1972) peuvent être utilisées comme signaux servant à la coordination de la parole. De par le caractère générique de notre modèle, l'interprétation de ces différents signaux multimodaux ne pose pas plus de difficultés théoriques que l'interprétation du volume sonore et de la hauteur de voix, tous les signaux étant traités comme des données variant continument de 0 à 1.

11.2.2 Adaptabilité de l'agent à l'utilisateur

Nous avons présenté dans le chapitre 7 une implémentation du modèle permettant une adaptation des agents à une absence de signaux non vocaux et à une interprétation bruitée des signaux du partenaire de l'agent. Néanmoins, la confrontation en temps réel de l'agent avec l'utilisateur montre une difficulté de l'agent à se coordonner avec l'utilisateur. Cette difficulté à se coordonner provient en grande partie d'imprécisions dans la détection de la hauteur de voix et du volume sonore de l'utilisateur. Dans ce cadre, le caractère adaptatif de l'agent ne semble pas suffire à compenser les imprécisions liées à la captation de la prosodie de l'utilisateur. Il est donc nécessaire d'améliorer la récupération des indices prosodiques dans un premier temps et d'évaluer de nouveau la capacité de l'agent à se coordonner avec un utilisateur. Il est aussi nécessaire, tel que mentionné au chapitre 10 d'apprendre un jeu de paramètre permettant de s'adapter à quelques utilisateurs. Nous comptons alors sur les capacités adaptatives offertes par les équations de perception et de contrôle de l'action de l'agent pour une adaptation plus générale à l'ensemble des utilisateurs dialoguant avec l'agent. Si cela s'avérait insuffisant, il serait nécessaire, à l'instar de Jonsdottir et Thórisson (2013), de proposer un mécanisme d'apprentissage modifiant les paramètres liés à la perception et au contrôle de l'action de l'agent. Néanmoins, la nature de notre modèle nous pousse à utiliser un autre mécanisme d'apprentissage

que celui proposé par Jonsdottir et Thórisson (2013) où l'agent apprend à éviter les recouvrements de parole et à prendre la parole sans moments de silence. En effet, nous devons tenir compte du fait que les situations liées au tour de parole sont variables, et qu'aucune des ces situations (transitions fluides, conflits de parole ou silences longs) n'est à privilégier par rapport à une autre.

Selon le paradigme de la dynamique comportementale (Warren, 2006), l'apprentissage se fait par exploration de l'espace des paramètres des lois de contrôle de l'agent, d'une manière analogue à un apprentissage par renforcement. L'agent modifie les paramètres de ses lois de contrôle et observe l'effet sur la progression de son comportement vers son but. Les lois de contrôle évoluent donc constamment, tant que l'agent n'a pas trouvé de solution comportementale, puis se stabilisent vers un ensemble de paramètres à mesure que l'état du système agent-utilisateur évolue vers le but de l'agent. Dans le cas de la coordination des tours de parole, l'agent a comme but de prendre la parole, de garder la parole ou de laisser la parole selon sa motivation à parler. Ces buts sont représentés par l'ensemble des attracteurs des signaux de l'agent : ainsi, prendre le tour implique un volume sonore et une hauteur de voix de l'agent moyens et un utilisateur ne parlant pas. Les lois de contrôle sont représentées par les équations de contrôle des signaux et les paramètres modifiés sont les paramètres définissant l'espace des attracteurs de l'agent. Aussi, l'agent modifie au cours du temps la fonction définissant l'attracteur de l'équation en fonction de l'information qu'il aura récolté sur la progression vers son but. S'il cherche à prendre la parole et que l'utilisateur ne souhaite pas la lui donner, il explore activement différentes variations de signaux pour essayer de pousser l'utilisateur à lui laisser la parole. Cela laisse comme problème principal le fait que l'agent est uniquement focalisé sur ses objectifs, sans se préoccuper des durées de conflit et des durées de silence qui doivent pourtant être limitées pour garantir qu'il y ait toujours une coordination entre les participants. Suivant les principes énoncés par De Loor *et al.* (2009), l'agent doit tenir compte à la fois de ses objectifs personnels mais aussi garder l'interaction « en vie » c'est-à-dire éviter un arrêt prématuré de l'interaction avec l'utilisateur. Cela doit être inclut dans les mécanismes d'apprentissage définissant l'évolution des paramètres de l'agent.

11.2.3 Traitement de l'information verbale

Au vu des expérimentations réalisées par De Ruiter *et al.* (2006) et Riest *et al.* (2015), il semble indéniable que l'interprétation de l'information verbale soit exploitée activement par les participants humains pour coordonner leurs tours. Pour la gestion du tour de parole utilisateur-agent l'information verbale contribuerait grandement à la détection de la fin de tour. Une telle capacité requiert pour un agent conversationnel une analyse en temps réel et incrémentale de l'énoncé de l'utilisateur en extrayant les informations sémantiques et syntaxiques de la phrase de l'utilisa-

teur. À notre connaissance, aucun agent actuel ne possède une telle capacité, les modèles traitant l'information verbale le faisant en dehors d'un contexte d'interaction en temps réel, sur des extraits annotés provenant d'interactions humaines (Huang *et al.*, 2011; De Vault *et al.*, 2011; Raux et Eskenazi, 2012). La question du traitement de l'information verbale est donc indéniablement une question de recherche clé dans le cadre de la gestion du tour de parole utilisateur-agent. De notre point de vue, pour avoir un agent crédible permettant une interaction naturelle avec l'utilisateur, la réponse à cette question doit provenir des modèles de psychologie sur le tour de parole. Or la question du traitement de l'information verbale par des participants humains est elle-même sujette à débat. Pour la grande majorité des auteurs sur le tour de parole, les participants prédisent le moment où la fin de tour du locuteur a lieu en détectant dans le discours de ce dernier des indices syntaxiques, grammaticaux et sémantiques Riest *et al.* (2015). Suivant ces principes De Vault *et al.* (2011) proposent un algorithme statistique permettant à un agent d'estimer la fin d'énoncé probable de l'agent. Cette approche par prédiction pose néanmoins question par les résultats expérimentaux montrant que les participants semblent en réalité avoir des difficultés à prédire les fins de phrase des utilisateurs tel que montré par Magyari et de Ruiter (2012). Si nous nous appuyons sur les principes d'une approche de perception-action continue, le traitement du contenu de l'énoncé pourrait être réalisé, de la même manière que pour l'information non-verbale, en accumulant de manière continue des indices verbaux informant de la progression de l'énoncé. Néanmoins la manière dont cette accumulation serait concrètement réalisée reste une question de recherche. Une autre problématique liée à l'implémentation des signaux verbaux réside dans la variation de la production verbale de l'agent liée aux différentes situations de coordination de la parole. Kurtić *et al.* (2013) mettent ainsi en avant la présence de « recyclages » lors des conflits : lorsqu'un conflit a lieu et que le locuteur ne souhaite pas laisser la parole, il a tendance à ne pas poursuivre son énoncé mais plutôt à répéter la dernière syllabe prononcée avant l'occurrence du conflit. Ces différentes expressions verbales pourraient être exploitées par un agent conversationnel pour améliorer la perception des différentes situations liées au tour de parole. L'emploi de synthèses vocales incrémentales rend possible la modification en temps réel de l'énoncé prononcé par l'agent permettant techniquement possible l'utilisation de recyclages. Néanmoins, notre modèle ne gère pour l'instant que la variation de grandeurs continues et la modulation de l'énoncé de l'agent nécessite une extension du modèle.

11.2.4 Extensions des fonctionnalités du modèle

Gestion multipartite du tour de parole

Nous avons proposé un modèle de coordination des tours de parole pour une interaction dyadique. Une extension possible de ce modèle réside dans son adaptation à des interactions multipartites, comportant plusieurs agents et plusieurs utilisateurs. Cette extension du modèle laisse plusieurs questions à résoudre. Une de ces problématiques concerne l'accumulation des indices de prise et fin de tour provenant de plusieurs participants. Dans les interactions humaines, le locuteur courant n'a pas de connaissance préalable du futur locuteur, il doit donc interpréter les signaux de tous ses auditeurs pour déterminer si un participant est en train de prendre le tour ou non. De plus, lorsqu'une prise de tour a lieu, il dirige son regard vers le participant prenant le tour. Cela implique que, non seulement le locuteur courant traque les signaux verbaux et non-verbaux des participants, mais qu'il est capable d'identifier le participant en train de prendre le tour. Notre modèle de tour de parole doit donc être en mesure d'identifier le comportement de chaque participant mais aussi d'associer une position spatiale à la valeur d'accumulation de chaque participant. Une solution serait d'implémenter plusieurs composantes d'accumulation interprétant le comportement de chaque partenaire de l'agent (agent ou utilisateur). Chaque composante d'accumulation serait associée avec la position spatiale du participant, et fonctionnerait en parallèle d'autres composantes d'accumulation. En sortie, les différentes valeurs d'accumulation seraient fusionnées et une donnée sur deux dimensions contenant la valeur de certitude concernant le participant prenant le tour et une information de direction du participant dans le repère de l'agent serait produite en sortie. L'information de direction varie en continu de la même manière que la valeur de certitude. Ainsi, si deux participants cherchent à prendre le tour en même temps, la valeur de direction pourra être située entre les deux participants montrant un agent incertain du participant actuel prenant le tour, ce qui pourrait être implémenté par un agent alternant les regards vers les deux participants. Une autre difficulté apparait pour l'auditeur lorsque l'on considère que le locuteur courant peut laisser la parole à un participant ou simplement finir son tour sans désigner quelqu'un en accord avec les observations effectuées par Sacks *et al.* (1974). Il est alors plus probable que le participant désigné par le locuteur courant comme prochain locuteur prenne la parole. Les auditeurs doivent donc non seulement déterminer si le locuteur est en train de laisser le tour mais plus particulièrement s'il désigne quelqu'un en particulier. Une solution serait de prendre en compte la direction du regard du locuteur, impactant positivement le calcul de la certitude s'il regarde l'agent et négativement le calcul de la certitude s'il ne regarde pas l'agent. Une autre problématique à résoudre lors du passage d'un modèle dyadique à un modèle multipartite concerne les différents rôles que peut prendre un auditeur. En effet, selon Clark (1996), un auditeur peut être soit interlocuteur (*addressee* en anglais),

c'est-à-dire un destinataire direct de l'énoncé du locuteur, un *side participant*, un auditeur reconnu par le locuteur courant comme participant à la conversation bien que n'étant pas un destinataire direct de l'énoncé du locuteur courant, ou un *overhearer*, n'étant pas considéré comme un participant de la conversation. Ces différents rôles ont une influence sur les échanges de parole. En effet, le locuteur courant s'attend avant tout à ce que le locuteur suivant soit l'un de ses interlocuteurs (Clark, 1996). Il est donc plus attentif aux signaux produits par ces participants qu'aux signaux produits par le *side participant*, et ne s'occupe pas du tout des signaux produits par les *overhearers*. De même, selon leur rôle, les auditeurs seront plus ou moins enclin à prendre la parole. Les interlocuteurs seront les moins réticents à prendre la parole tandis que le *side participant* ne prendra la parole que s'il juge l'information suffisamment importante pour outrepasser le rôle qui lui est attribué, les *overhearers* ne prenant pas la parole. Si nous nous intéressons uniquement aux comportements des interlocuteurs et des *side participants* dans notre modèle, nous devons ajouter une variable de contrôle des signaux de l'agent calculant le fait que le locuteur courant s'adresse à l'agent en particulier. Cette valeur est, à l'instar de la certitude sur la fin de tour du locuteur courant, continue et accumulée au cours du temps par un processus d'accumulation fonctionnant en parallèle des autres processus.

Ajout de comportements d'écoute

La gestion des comportements d'écoute de l'auditeur, c'est-à-dire la capacité à produire au bon moment des *backchannels* pour l'auditeur (Bevacqua *et al.*, 2008) et à interpréter ces *backchannels* et y réagir pour le locuteur (Reidsma *et al.*, 2011), est une thématique proche de celle de la gestion du tour de parole, et constitue une extension intéressante de notre modèle de coordination de la parole. Le déclenchement des *backchannels* serait géré par une équation de contrôle pondérant l'intensité du *backchannel* produit par le participant. L'auditeur utiliserait alors un autre processus d'accumulation, interprétant les signaux d'invitation au *backchannel* du locuteur courant comme variable de contrôle de l'équation de contrôle. La valeur de motivation de l'agent ne serait plus seulement entre parler et ne pas parler mais entre parler et écouter, et constituerait une seconde variable de contrôle de la production de *backchannels* de l'auditeur. Selon Gravano et Hirschberg (2011), le locuteur module son volume sonore et sa hauteur de voix pour inviter l'auditeur à produire un *backchannel*. Nous devons donc inclure une variable de contrôle dans les équations de contrôle du volume sonore et de la hauteur de voix renseignant une forme de nécessité à avoir le retour de l'auditeur. Cette variable serait le résultat d'un processus d'accumulation surveillant les *backchannels* de l'auditeur. D'une manière similaire à Buschmeier et Kopp (2014), lorsque ce dernier produit un *backchannel*, la nécessité du retour de l'auditeur diminue à la valeur minimale tandis que l'absence de *backchannels* fait croitre au cours du temps cette variable. L'approche théorique prise

au cours de cette thèse permet tout à fait l'implémentation de ces comportements d'écoute.

Annexe A

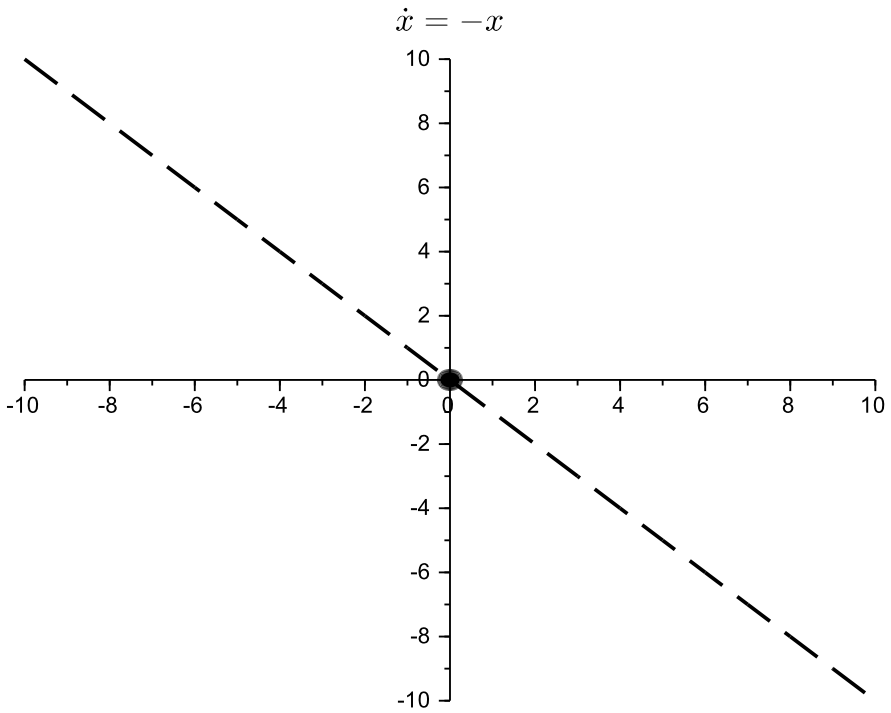
Systèmes dynamiques

Warren (2006) propose une bonne introduction à la modélisation des systèmes dynamiques, nous reprenons ses explications ici.

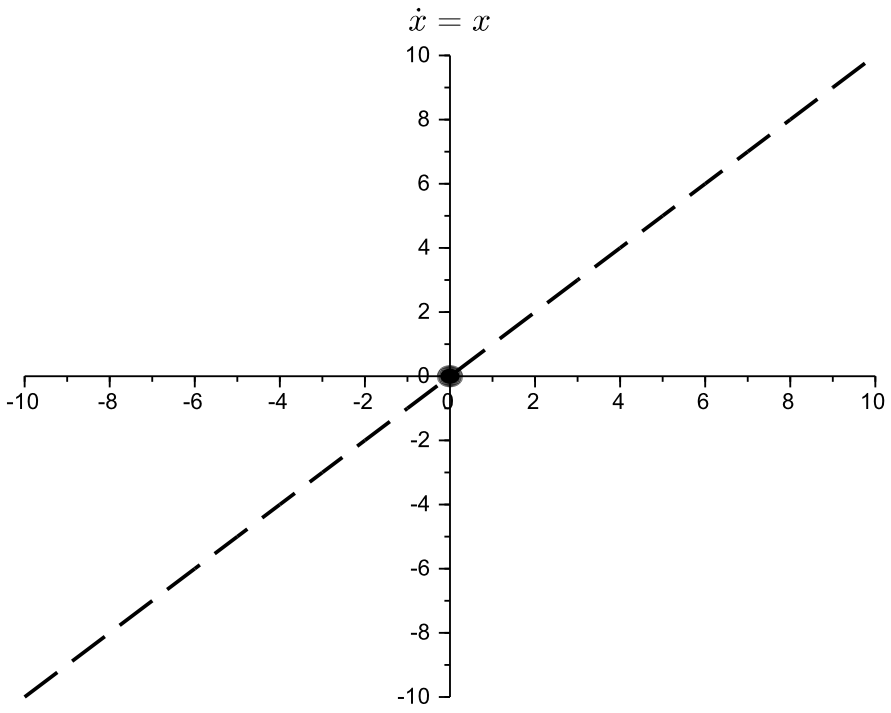
Les systèmes dynamiques sont des outils mathématiques modélisant le comportement de systèmes divers sous forme d'équations différentielles. Ces équations différentielles spécifient l'évolution de la dérivée (variation) d'une variable par rapport à la valeur courante de cette variable. La formule donnée par cette équation différentielle spécifie un espace d'état définissant, lui, l'évolution des valeurs de la variable au cours du temps en fonction des valeurs initiales. L'espace d'état est un autre mode de représentation d'une équation différentielle où l'équation différentielle est traitée comme une fonction $\dot{x} = f(x)$. Prenons l'exemple d'une équation simple : $\dot{x} = -x$. L'espace d'état représente en conséquence la dérivée de x en fonction de x. Cette équation donne lieu à l'espace d'état donné sur la figure A.1a. Peu importe la position des conditions initiales ce système convergera vers la valeur 0. En effet lorsque la valeur initiale du système est positive, la dérivée correspondante est négative, impliquant que l'état du système décroit au fil du temps. Plus x diminue plus sa valeur de dérivée diminue vers 0. Lorsque le point est proche de 0, la valeur de dérivée est quasi-nulle. x se rapproche alors de manière asymptotique vers la valeur 0. De même si les conditions initiales sont négatives, la dérivée est positive et l'état du système évoluera aussi vers l'origine. On a affaire ici à un point stable à l'origine, si l'état du système est à 0, une perturbation dans le système sera rapidement compensée et l'état du système convergera de nouveau vers 0.

À l'inverse, si l'on prend l'équation $\dot{x} = x$, on obtient l'espace d'état montré sur la figure A.1b. Dans cette configuration, le système est instable et soit le système diverge vers $-\inf$, soit le système diverge vers $+\inf$. Toute perturbation du système engendrera une déstabilisation du système. En ce sens le point fixe représenté en $(0, 0)$ est un répulseur du système.

Certaines équations proposent d'autres formes de points fixes. L'équation $\dot{x} = -x^2$ possède par exemple un point fixe semi-stable à l'origine du repère. La représentation graphique de cet espace d'état est montré sur la figure A.2. Lorsque l'état



(a) Espace d'état du système $\dot{x} = -x$ le point en coordonnées (0,0) constitue l'attracteur du système.



(b) Espace d'état du système $\dot{x} = x$ le point en coordonnées (0,0) constitue un répulseur du système.

FIGURE A.1 – Espace d'états respectivement des équations $\dot{x} = -x$ (à gauche) et $\dot{x} = x$ (à droite).

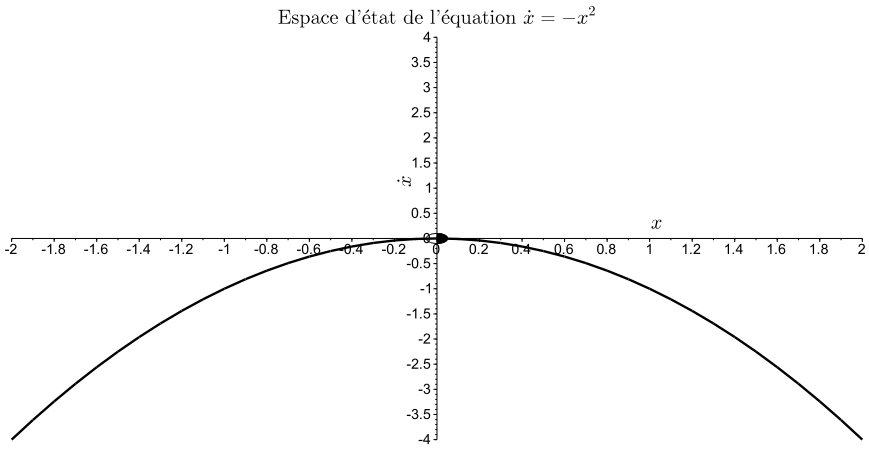


FIGURE A.2 – Équation non-linéaire $\dot{x} = -x^2$ donnant un point fixe semi-stable.

initial du système est supérieur à 0 le système converge vers 0. Néanmoins, lorsque l'état du système est négatif, le système est repoussé vers $-\inf$.

Considérons maintenant le cas d'une équation différentielle du second ordre. L'une des plus simples est définie par l'équation A.1.

$$\ddot{x} = -b \times \dot{x} - k_g \times x \tag{A.1}$$

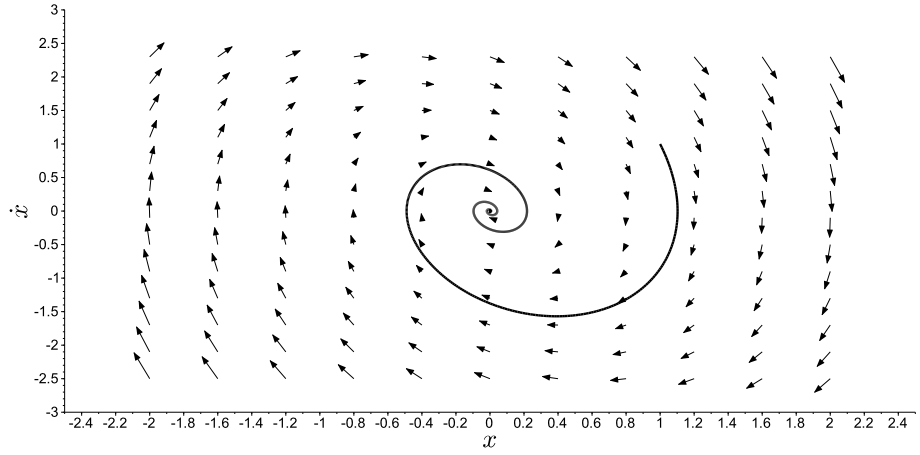
Cette équation représente la forme générale d'équations de systèmes amortis. En physique, cette équation a été utilisée pour modéliser le comportement de systèmes physiques comme le comportement d'un ressort avec une masse à son bout, le comportement d'un pendule.

En fonction de la relation entre b et k_g, on observe trois catégories de systèmes montrés sur la figure A.3. Si $k_g \geq \frac{b^2}{4}$, on a un système pseudo-périodique. Dans ce cas de figure, le système possède un point fixe attracteur en 0, et le système oscille autour de l'attracteur en convergeant vers lui (figure A.3a). Si $k_g < \frac{b^2}{4}$, on a un système apériodique, le système possède dans ce cas là un attracteur en 0 et le système possède une asymptote en 0 de sorte que x ne change jamais de signe (figure A.3b). Un cas particulier est défini lorsque $b = 0$, dans ce cas là, l'attracteur du système n'est plus un point fixe mais un cycle limite. Le système converge vers une trajectoire asymptotique dans l'espace d'état oscillant en permanence autour de l'origine du repère (figure A.3c).

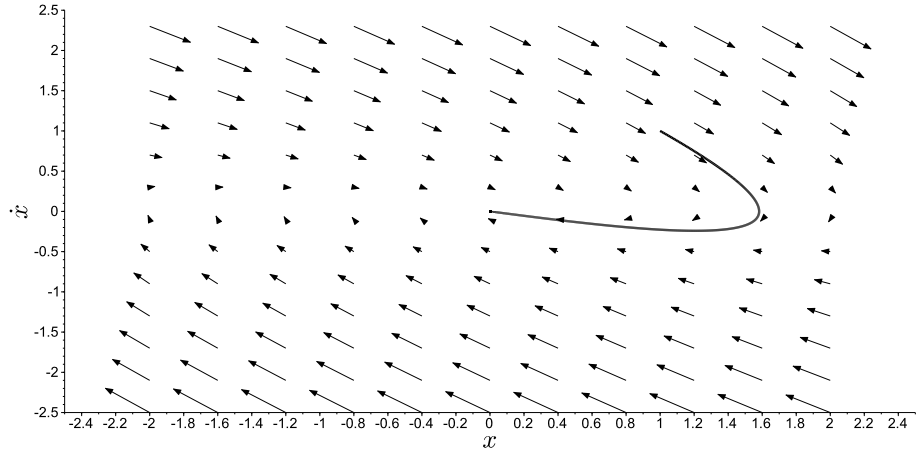
Si nous ajoutons maintenant des paramètres variant au cours du temps, nous modifions la disposition de l'espace d'état. Ajoutons par exemple le paramètre $a(t)$ de sorte que l'équation différentielle soit définie par l'équation A.2.

$$\ddot{x} = -b \times a(t) \times \dot{x} - k_g \times x \tag{A.2}$$

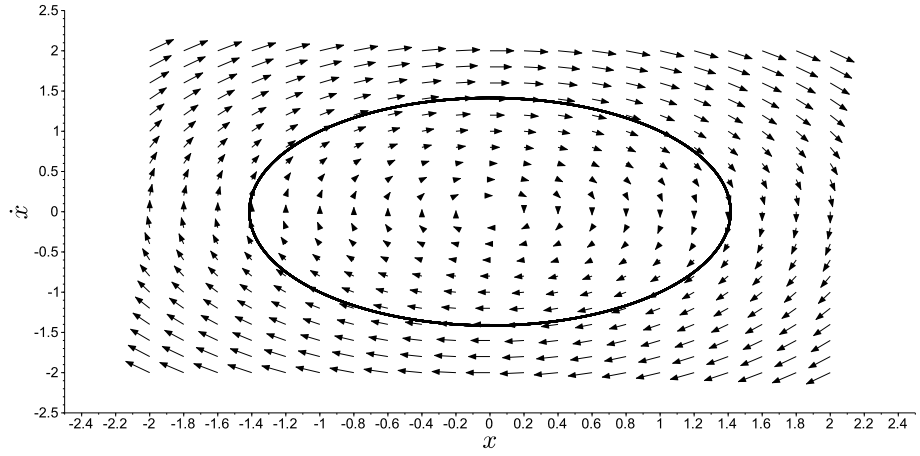
Imaginons que $a(t)$ varie linéairement en fonction du temps. Si l'on considère



(a) Système en régime pseudo-périodique. Le système a un point fixe attracteur, et oscille autour de l'attracteur avant de le rejoindre.



(b) Système en régime apériodique. Le système a un point fixe attracteur et converge vers cet attracteur sans osciller autour.



(c) Système en oscillation stable. Le système a comme attracteur un cycle limite.

FIGURE A.3 – Trois types d'espace d'état définis par l'équation $\ddot{x} = -b \times \dot{x} - k_g \times x$.

$a(0) = 0$, on obtient un oscillateur parfait avec un cycle limite comme attracteur tel que défini sur la figure A.3c. À mesure de $a(t)$ augmente, tant que $k_g \geq \frac{b^2 \times a(t_i)^2}{4}$, le système a un régime pseudo-périodique, puis devient apériodique lorsque $k_g < \frac{b^2 \times a(t_i)^2}{4}$. Une autre manière de provoquer une bifurcation est de rajouter un terme $a(t)$ à l'équation donnant l'équation A.3.

$$\ddot{x} = -b \times \dot{x} - k_g \times x + a(t) \tag{A.3}$$

Imaginons que $b > 0$, le système possède alors un point fixe. À mesure que $a(t)$ varie le point fixe du système varie vers la valeur $a(t)$ et la trajectoire du système convergera vers ce point fixe.

Annexe B

Modèle de dérive-diffusion (*Drift Diffusion Model* ou DDM)

L'équation du DDM se présente sous la forme donnée par l'équation B.1.

$$dx = Adt + cdW \tag{B.1}$$

Il s'agit d'une équation stochastique représentant un processus aléatoire. dx représente la variation d'accumulation d'indices, A représente le taux d'accroissement moyen de la différence dans les indices accumulés, cdW est un terme aléatoire suivant une loi normale centrée, de variance c. Le DDM définit de plus deux seuils de décision, un seuil positif et un seuil négatif représentant les deux alternatives possibles. Si le taux d'accumulation est négatif, la quantité d'indices variera en moyenne vers le seuil négatif, si le taux d'accumulation est positif, la quantité d'indices variera en moyenne vers le seuil positif. Ces deux seuils peuvent être égaux en matière de valeur absolue, ou différents. Une différence représente un biais dans la prise de décision, l'agent est au début du processus de décision plutôt favorable envers l'alternative dont le seuil a la valeur absolue la plus petite. La quantité d'indices de départ de l'agent peut être soit nulle, soit non nulle. Une quantité d'indices non nulle représente, de la même manière qu'une différence dans les seuils, un biais dans la décision.

Selon les applications du modèle, le taux d'accumulation peut être soit constant pendant le processus de prise de décision, soit variable (Ratcliff, 1980). Le choix d'un A constant s'effectue lorsque l'environnement de l'agent ne varie pas : les informations reçues par l'agent ne changent pas. Au contraire dans un environnement dynamique, les indices reçus varient au cours du temps, changeant le taux d'accumulation. Un exemple de simulation selon le *DDM* est montré sur la figure B.1. Dans cet exemple de simulation, nous avons fixé un paramètre stochastique avec un écart-type $c = 0.4$. Le paramètre de dérive A est nul de 0 à 3 secondes puis devient positif après trois secondes. La figure montre une croissance dans la variation de

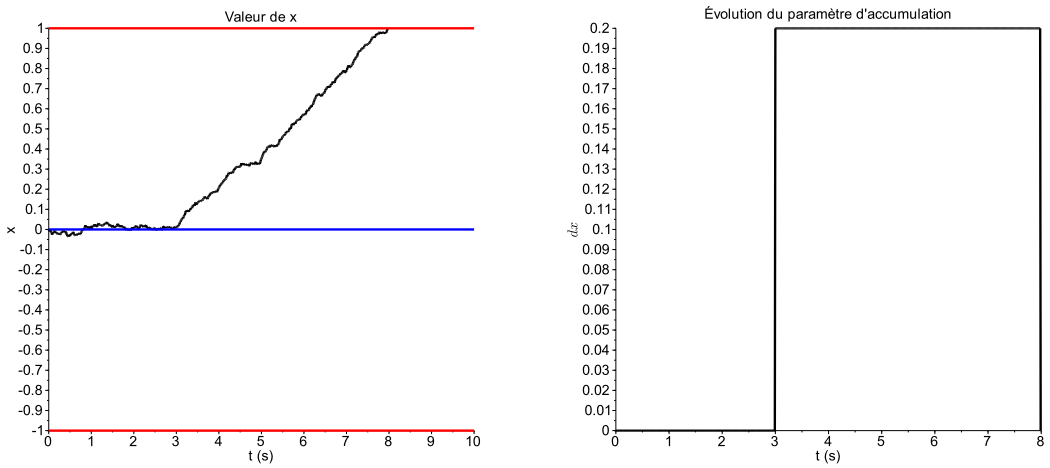


FIGURE B.1 – Exemple d'un processus de prise de décision modélisé par l'équation du *DDM*, les seuils de décisions sont indiqués en 1 et -1 et le biais de la décision en 0.

quantité d'indices lorsque le paramètre A devient non nul. La variation de la quantité d'indices devient positive jusqu'à atteindre le seuil de décision positif. Arrivé au seuil de décision, la simulation s'arrête.

Annexe C

Détails du paramétrage du modèle

Le scénario de coordination de la parole sur plusieurs tours a été réalisé en modifiant la motivation à parler pour les deux agents selon les machines à états présentées figure 9.10. Le modèle a été paramétré avec les fonctions d'accumulation partielles présentées dans le tableau C.1.

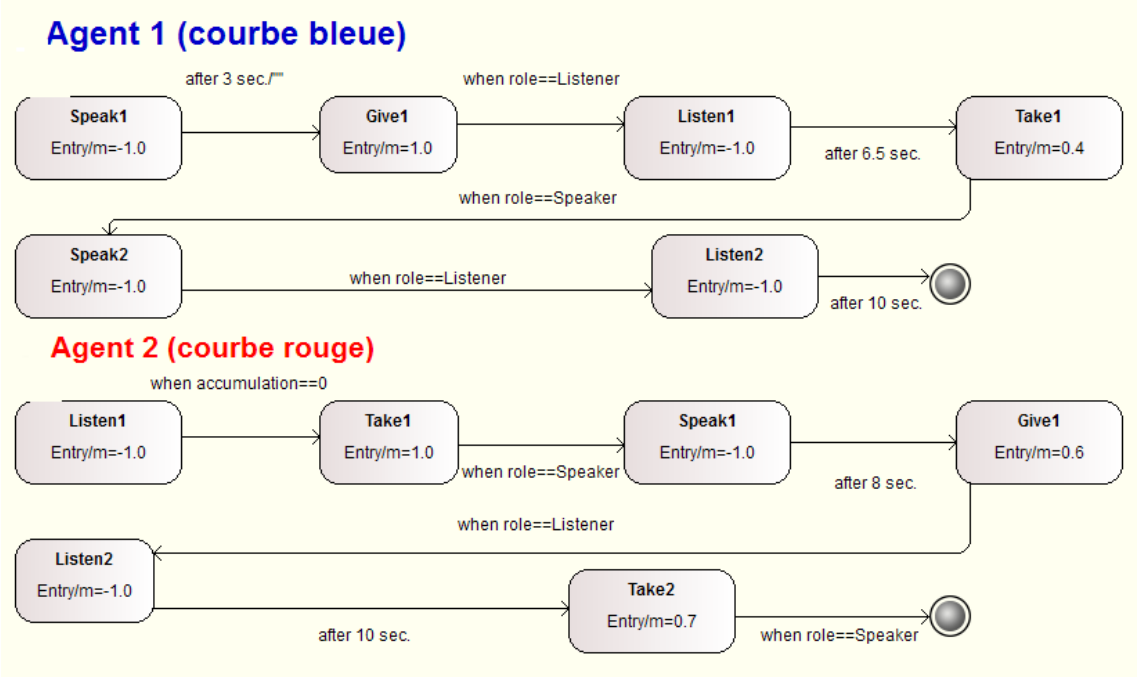


FIGURE C.1 – Machines à états utilisées pour contrôler les motivations à changer de rôle des deux participants dans le cadre du scénario de la figure 9.10

	Locuteur	Auditeur
Hauteur de voix	$b_0 = -0.3$, $b_1 = 1.8$, $b_2 = 0.3$	$b_0 = 0.45$, $b_1 = -1.0$, $b_2 = -0.2$
Volume sonore	$b_0 = -0.5$, $b_1 = 1.8$, $b_2 = 0.3$	$b_0 = 0.45$, $b_1 = -1.0$, $b_2 = -0.2$

TABLE C.1 – Équations d'accumulation partielle des composantes de perception du comportement de l'agent.

Bibliographie

Al Moubayed, S. et Lehman, J. (2015). Regulating Turn-Taking in Multi-child Spoken Interaction. *In* Brinkman, W.-P., Broekens, J. et Heylen, D., éditeurs : *Intelligent Virtual Agents*, numéro 9238 de Lecture Notes in Computer Science, pages 363–374. Springer International Publishing. DOI : 10.1007/978-3-319-21996-7_40.

Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA, USA : Harvard University Press.

André, E., Dorfmüller-Ulhaas, K. et Rehm, M. (2005). Engaging in a converation with synthetic characters : along the virtuality continuum. *In 5th International Symposium, SG 2005, Frauenwörth Cloister, Germany, August 22-24, 2005. Proceedings*, pages 1–12.

Aponte, M.-V., Levieux, G. et Natkin, S. (2011). Measuring the level of difficulty in single player video games. *Entertainment Computing*, 2(4):205–213.

Axelrod, R. (1981). *The evolution of cooperation*. New York : Basic Books.

Bailenson, J. N. et Yee, N. (2005). Digital chameleons automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological science*, 16(10):814–819.

Bailly, G. et Gouvernayre, C. (2012). Pauses and respiratory markers of the structure of book reading. *In 13th Annual Conference of the International Speech Communication Association (InterSpeech 2012)*, page Thu.O9d.05, Portland, United States.

Balentine, B. E., Ayer, C. M., Miller, C. L. et Scott, B. L. (1997). Debouncing the speech button : A sliding capture window device for synchronizing turn-taking. *International Journal of Speech Technology*, 2(1):7–19.

Baumann, T. (2013). *Incremental spoken dialogue processing : Architecture and lower-level components*. Thèse de doctorat, Université de Bielefeld.

Bavelas, J. B., Chovil, N., Coates, L. et Roe, L. (1995). Gestures Specialized for Dialogue. *Personality and Social Psychology Bulletin*, 21(4):394–405.

Bernstein, B. (1962). Linguistic Codes, Hesitation Phenomena and Intelligence. *Language and Speech*, 5(1):31–48.

Berry, A. (1994). Spanish and American Turn-Taking Styles : A Comparative Study. *Pragmatics and Language Learning*, 5:180–194.

Berry, D. C., Butler, L. T. et de Rosis, F. (2005). Evaluating a realistic agent in an advice-giving task. *International Journal of Human-Computer Studies*, 63(3): 304–327.

Beňuš, t., Gravano, A. et Hirschberg, J. (2011). Pragmatic aspects of temporal accommodation in turn-taking. *Journal of Pragmatics*, 43(12):3001–3027.

Bevacqua, E., Mancini, M. et Pelachaud, C. (2008). A Listening Agent Exhibiting Variable Behaviour. *In* Prendinger, H., Lester, J. et Ishizuka, M., éditeurs : *Intelligent Virtual Agents*, pages 262–269. Springer Berlin Heidelberg.

Bevacqua, E., Pammi, S., Hyniewska, S. J., Schröder, M. et Pelachaud, C. (2010). Multimodal backchannels for embodied conversational agents. *In 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings*, pages 194–200.

Bevacqua, E., Prepin, K., de Sevin, E., Niewiadomski, R. et Pelachaud, C. (2009). Reactive behaviors in SAIBA architecture. *In Workshop on Towards a Standars Markup Language for Embodied Dialogue Acts. Autonomous Agents and Multi-Agent Systems*, pages 9–12, Budapest, Hongrie.

Bevacqua, E., Stanković, I., Maatallaoui, A., Nédélec, A. et De Loor, P. (2014). Effects of Coupling in Human-Virtual Agent Body Interaction. *In Intelligent Virtual Agents 2014*, pages 54–63, Boston, MA.

Bögels, S., Magyari, L. et Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports*, 5:12881.

Bögels, S. et Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52:46–57.

Bicchieri, C. (1989). Self-refuting theories of strategic interaction : A paradox of common knowledge. *Erkenntnis*, 30(1-2):69–85.

Bickmore, T., Pfeifer, L. et Schulman, D. (2011). Relational Agents Improve Engagement and Learning in Science Museum Visitors. *In Proceedings of Intelligent Virtual Agent 2011*, pages 55–67.

Bickmore, T., Schulman, D. et Yin, L. (2010). Maintaining engagement in long-term interventions with relational agents. *Applied Artificial Intelligence*, 24(6): 648–666.

Bickmore, T. W. et Picard, R. W. (2005). Establishing and Maintaining Long-term Human-computer Relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):293–327.

Binmore, K. (1987). Modeling Rational Players : Part I. *Economics and Philosophy*, 3(2):179–214.

Biocca, F. (1997). The Cyborg's Dilemma : Progressive Embodiment in Virtual Environments. *Journal of Computer Mediated Communication*, 3(2):0.

Biocca, F., Harms, C. et Burgoon, J. K. (2003). Towards A More Robust Theory and Measure of Social Presence : Review and Suggested Criteria. *Presence : Teleoperators and Virtual Environments*, 12(5):456–480.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot International*, 5(9):341–345.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P. et Cohen, J. (2006). The physics of optimal decision making : a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, 113(4):700.

Bohus, D. et Horvitz, E. (2011). Decisions About Turns in Multiparty Conversation : From Perception to Action. *In Proceedings of the 13th International Conference on Multimodal Interfaces*, pages 153–160.

Boyle, E. A., Connolly, T. M., Hainey, T. et Boyle, J. M. (2012). Engagement in digital entertainment games : A systematic review. *Computers in Human Behavior*, 28(3):771 – 780.

Brams, S. J. (1975). *Game Theory and Politics*. Mineola,NY : Dover Publications.

Brockmyer, J. H., Fox, C. M., Curtiss, K. A., McBroom, E., Burkhart, K. M. et Pidruzny, J. N. (2009). The development of the Game Engagement Questionnaire : a measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, 45(4):624–634.

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal on Robotics and Automation*, 2(1):14–23.

Burgoon, J. K., Bonito, J. A., Bengtsson, B., Cederberg, C., Lundeberg, M. et Allspach, L. (2000). Interactivity in human–computer interaction : a study of credibility, understanding, and influence. *Computers in Human Behavior*, 16(6):553–574.

Buschmeier, H. et Kopp, S. (2014). When to elicit feedback in dialogue : Towards a model based on the information needs of speakers. *In Proceedings of Intelligent Virtual Agents*, pages 71–80, Boston,MA.

Caelen, J. (2003). Stratégies de dialogue. *In MFI'03 (Modèles formels de l'interaction)*, Lille.

Cafaro, A., Vilhjálmsson, H. H., Bickmore, T., Heylen, D. et Pelachaud, C. (2014). Representing Communicative Functions in SAIBA with a Unified Function Markup Language. *In Intelligent Virtual Agents*, pages 81–94. Springer.

Campione, E. et Véronis, J. (2002). A Large-Scale Multilingual Study of Silent Pause Duration. *In ESCA-workshop on speech prosody*, pages 199–202, Aix-en-Provence.

Casillas, M., Bobb, S. C. et Clark, E. V. (2015). Turn-taking, timing, and planning in early language acquisition. *Journal of Child Language*, First View:1–28.

Cassel, J. (2000). More than just another pretty face : embodied conversational interface agents. *Communications of the ACM*, 43(4):70–78.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H. et Yan, H. (1999). Embodiment in conversational interfaces. *In Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 520–527.

Chapman, P., Selvarajah, S. et Webster, J. (1999). Engagement in multimedia training systems. *In Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences, 1999. HICSS-32*, pages 1084–1093, Maui, Hawai, USA.

Chowdhury, S. A., Danieli, M. et Riccardi, G. (2015). Annotating and categorizing competition in overlap speech. *In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5316–5320. IEEE.

Clancy, B. et McCarthy, M. (2015). Co-constructed turn-taking. *In* Aijmer, K. et Rühlemann, C., éditeurs : *Corpus Pragmatics*, pages 430–453. Cambridge University Press, Cambridge.

Clark, H. H. (1996). *Using Language*. Cambridge, England : Cambridge University Press.

Clark, H. H. et Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and brain sciences*, 26(02):139–153.

Cummins, F. (2012). Oscillators and syllables : A cautionary note. *Frontiers in Psychology*, 3(364).

Cutler, A. et Pearson, M. (1986). On the analysis of prosodic turn-taking cues. *Intonation in discourse*, pages 139–156.

De Greef, P. et Ijsselsteijn, W. A. (2000). Social Presence in a Home Tele-Application. *CyberPsychology & Behavior*, 4(2):307–315.

de Kok, I. A. (2013). *Listening heads*. Thèse de doctorat, University of Twente, Enschede.

De Loor, P., Bevacqua, E., Stankovic, I., Maatallaoui, A., Nédélec, A. et Buche, C. (2015). Le couplage d'agents virtuels interactifs socialement présents. *Vers une communication Homme-Animal-Machine ? : Contribution interdisciplinaire*, pages 237–254.

De Loor, P., Manac'h, K. et Tisseau, J. (2009). Enaction-Based Artificial Intelligence : Toward Co-evolution with Humans in the Loop. *Minds and Machines*, 19(3):319–343.

De Ruiter, J. P., Mitterer, H. et Enfield, N. J. (2006). Projecting the end of a speaker's turn : A cognitive cornerstone of conversation. *Language*, 82(3):515–535.

De Vault, D., Mell, J. et Gratch, J. (2015). Toward natural turn-taking in a virtual human negotiation agent. *In AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction*, Stanford, CA.

De Vault, D., Sagae, K. et Traum, D. (2011). Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue & Discourse*, 2(1):143–170.

Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S. et Cohen, D. (2012). Interpersonal synchrony : A survey of evaluation methods across disciplines. *Affective Computing, IEEE Transactions on*, 3(3):349–365.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292.

Duncan, S. et Niederehe, G. (1974). On signalling that it's your turn to speak. *Journal of Experimental Social Psychology*, 10(3):234–247.

Elliott, C. et Brzezinski, J. (1998). Autonomous agents as synthetic characters. *AI magazine*, 19(2):13.

Ellis, C. A., Gibbs, S. J. et Rein, G. (1991). Groupware : Some Issues and Experiences. *Communications of ACM*, 34(1):39–58.

Eyben, F., Weninger, F., Gross, F. et Schuller, B. (2013). Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. *In Proceedings of the 21st ACM International Conference on Multimedia*, pages 835–838.

Fajen, B. R. (2013). Guiding locomotion in complex dynamic environments. *Frontiers in Behavioral Neuroscience*, 7(85).

Fajen, B. R. et Warren, W. H. (2007). Behavioral dynamics of intercepting a moving target. *Experimental Brain Research*, 180(2):303–319.

Fowler, C. A., Richardson, M. J., Marsh, K. L. et Shockley, K. D. (2008). Language Use, Coordination, and the Emergence of Cooperative Action. *In* Fuchs, A. et Jirsa, V. K., éditeurs : *Coordination : Neural, Behavioral and Social Dynamics*, pages 261–279. Springer Berlin Heidelberg.

Fox Tree, J. E. (2000). Coordinating spontaneous talk. *Aspects of Language Production*, pages 375–406.

Frank, T. D., Richardson, M. J., Lopresti-Goodman, S. M. et Turvey, M. T. (2009). Order Parameter Dynamics of Body-scaled Hysteresis and Mode Transitions in Grasping Behavior. *Journal of Biological Physics*, 35(2):127–147.

Fuks, H., Raposo, A. et Gerosa, M. A. (2007). The 3c collaboration model. *In* (org), N. K., éditeur : *Encyclopedia of E-Collaboration*, pages 637–643. Hershey, PA : Information Science Reference.

Galantucci, B. et Sebanz, N. (2009). Joint Action : Current Perspectives. *Topics in Cognitive Science*, 1(2):255–259.

Gibson, J. J. (1979). *The Ecological Approach To Visual Perception*. New York, NY : Psychology Press.

Giles, H. et Coupland, N. (1991). *Language : Contexts and Consequences*. Belmont, CA : Open University Press.

Goffman, E. (1976). Replies and responses. *Language in Society*, 5(03):257–313.

Goodwin, C. (1981). *Conversational organization : Interaction between speakers and hearers*. New York, NY : Academic Press.

Gravano, A. et Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

Groom, V., Nass, C., Chen, T., Nielsen, A., Scarborough, J. K. et Robles, E. (2009). Evaluating the effects of behavioral realism in embodied agents. *International Journal of Human-Computer Studies*, 67(10):842–849.

Grosjean, F. et Hirt, C. (1996). Using Prosody to Predict the End of Sentences in English and French : Normal and Brain-Damaged Subjects. *Language and Cognitive Processes*, 11(1-2):107–134.

Haken, H., Kelso, J. a. S. et Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*, 51(5):347–356.

Hartholt, A., Traum, D., Marsella, S. C., Shapiro, A., Stratou, G., Leuski, A., Morency, L.-P. et Gratch, J. (2014). All Together Now : Introducing the Virtual Human Toolkit. *In Workshop on Architectures and Standards for Intelligent Virtual Agents*, Boston,MA.

Hayes-Roth, B. et Doyle, P. (1998). Animate Characters. *Autonomous Agents and Multi-Agent Systems*, 1(2):195–230.

Hecht, M. A. et Ambady, N. (1999). Nonverbal communication and psychology : Past and future. *Atlantic Journal of Communication*, 7(2):156–170.

Heldner, M. et Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, 53(1):23–35.

Holler, J. et Kendrick, K. H. (2015). Unaddressed participants' gaze in multi-person interaction : optimizing recipiency. *Frontiers in Psychology*, 6(98).

Huang, L., Morency, L.-P. et Gratch, J. (2011). A Multimodal End-of-turn Prediction Model : Learning from Parasocial Consensus Sampling. *In The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 3*, AAMAS '11, pages 1289–1290, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Ikegami, T. et Iizuka, H. (2007). Turn-taking Interaction as a Cooperative and Co-creative Process. *Infant Behavior and Development*, 30(2):278–288.

Ishii, R., Miyajima, T., Fujita, K. et Nakano, Y. (2006). Avatar's Gaze Control to Facilitate Conversational Turn-Taking in Virtual-Space Multi-user Voice Chat System. *In* Gratch, J., Young, M., Aylett, R., Ballin, D. et Olivier, P., éditeurs : *Intelligent Virtual Agents*, numéro 4133 de Lecture Notes in Computer Science, pages 458–458. Springer Berlin Heidelberg.

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. *Conversation analysis : studies from the first generation*, 125:13–31.

Johansson, M., Skantze, G. et Gustafson, J. (2014). Comparison of Human-Human and Human-Robot Turn-Taking Behaviour in Multiparty Situated Interaction. *In Proceedings of the 2014 Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, UM3I '14, pages 21–26, New York, NY, USA. ACM.

Jonsdottir, G. R., Thorisson, K. R. et Nivel, E. (2008). Learning Smooth, Human-Like Turntaking in Realtime Dialogue. *In* Prendinger, H., Lester, J. et Ishizuka, M., éditeurs : *Intelligent Virtual Agents*, numéro 5208 de Lecture Notes in Computer Science, pages 162–175. Springer Berlin Heidelberg.

Jonsdottir, G. R. et Thórisson, K. R. (2013). A Distributed Architecture for Real-time Dialogue and On-task Learning of Efficient Co-operative Turn-taking. *In* Campbell, N., éditeur : *Coverbal Synchrony in Human-Machine Interaction*, pages 293–323.

Jump, A. et Ekholm, J. (2015). Virtual Personal Assistant Use Is Growing, but Usage Functions Are Still Limited. https ://www.gartner.com/doc/2956117/virtual-personal-assistant-use-growing.

Keitel, A. et Daum, M. M. (2015). The use of intonation for turn anticipation in observed conversations without visual signals as source of information. *Frontiers in Psychology*, 6(108).

Kelso, J. (2009). Coordination dynamics. *Encyclopedia of complexity and systems sciences*, pages 1537–1564.

Kelso, J. A. S., de Guzman, G. C., Reveley, C. et Tognoli, E. (2009). Virtual Partner Interaction (VPI) : Exploring Novel Behaviors via Coordination Dynamics. *PLoS ONE*, 4(6):e5749.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63.

Kilpatrick, P. (1986). Turn and Control in Puerto Rican Spanish Conversation. Research/Technical ED269983, University of Puerto Rico, Mayaguez, Puerto Rico.

Kopp, S., Krenn, B., Marsella, S., Marshall, A. N., Pelachaud, C., Pirker, H., Thórisson, K. R. et Vilhjálmsson, H. (2006). Towards a common framework for multimodal generation : The behavior markup language. *In IVA'06 Proceedings of the 6th international conference on Intelligent Virtual Agents*, pages 205–217.

Kopp, S., Welbergen, H. v., Yaghoubzadeh, R. et Buschmeier, H. (2014). An architecture for fluid real-time conversational agents : integrating incremental output generation and input processing. *Journal on Multimodal User Interfaces*, 8(1):97–108.

Kurtić, E., Brown, G. J. et Wells, B. (2013). Resources for turn competition in overlapping talk. *Speech Communication*, 55(5):721–743.

Laird, J. E., Newell, A. et Rosenbloom, P. S. (1987). Soar : An Architecture for General Intelligence. *Artificial Intelligence*, 33(1):1–64.

Lammertink, I., Casillas, M., Benders, T., Post, B. et Fikkert, P. (2015). Dutch and English toddlers' use of linguistic cues in predicting upcoming turn transitions. *Frontiers in Psychology*, 6(495).

Lee, K. M. et Nass, C. (2003). Designing social presence of social actors in human computer interaction. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 289–296. ACM Press.

Lepora, N. F. et Pezzulo, G. (2015). Embodied Choice : How Action Influences Perceptual Decision Making. *PLOS Computational Biology*, 11(4):e1004110.

Lessmann, N., Kranstedt, A. et Wachsmuth, I. (2004). Towards a cognitively motivated processing of turn-taking signals for the embodied conversational agent Max. *In Proceedings of the Workshop Embodied Conversational Agents : Balanced Perception and Action*, page 65, New-York, August 19–23. ACM Press.

Levine, J. M. et Resnick, L. B. (1993). Social Foundations Of Cognition. *Annual review of psychology*, 44(1):585–612.

Lewis, D. (1969). *Convention : A Philosophical Study*. Cambridge, MA : Harvard University Press.

Lombard, M. et Ditton, T. (1997). At the Heart of It All : The Concept of Presence. *Journal of Computer Mediated Communication*, 3(2):0–0.

Loomis, J. M. et Beall, A. C. (2004). Model-based control of perception/action. *In Optic flow and beyond*, pages 421–441. Springer.

Lucas, G. M., Gratch, J., King, A. et Morency, L.-P. (2014). It's only a computer : Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100.

Magyari, L. et de Ruiter, J. P. (2012). Prediction of Turn-Ends Based on Anticipation of Upcoming Words. *Frontiers in Psychology*, 3(376).

Marsh, K. L., Richardson, M. J., Baron, R. M. et Schmidt, R. (2006). Contrasting Approaches to Perceiving and Acting With Others. *Ecological Psychology*, 18(1):1–38.

Maturana, H. R. et Varela, F. J. (1980). *Autopoiesis and Cognition*, volume 42 de *Boston Studies in the Philosophy of Science*. Dordrecht : Springer Netherlands.

McFarland, D. H. (2001). Respiratory markers of conversational interaction. *Journal of Speech, Language, and Hearing Research*, 44:128–143.

Niebuhr, O., Görs, K. et Graupe, E. (2013). Speech Reduction, Intensity, and F0 Shape are Cues to Turn-Taking. *In Proceedings of the SIGDIAL 2013 Conference*, pages 261–269, Metz.

Niewiadomski, R., Demeure, V. et Pelachaud, C. (2010). Warmth, Competence, Believability and Virtual Agents. *In* Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C. et Safonova, A., éditeurs : *Intelligent Virtual Agents*, numéro 6356 de Lecture Notes in Computer Science, pages 272–285. Springer Berlin Heidelberg. DOI : 10.1007/978-3-642-15892-6_29.

Nooraei, B., Rich, C. et Sidner, C. (2014). A Real-Time Architecture for Embodied Conversational Agents : Beyond Turn-Taking. *In ACHI 2014, The Seventh International Conference on Advances in Computer-Human Interactions*, pages 381–388.

Novick, D., Hansen, B. et Ward, K. (1996). Coordinating turn-taking with gaze. *In Fourth International Conference on Spoken Language, 1996. ICSLP 96. Proceedings*, volume 3, pages 1888–1891 vol.3.

O'Brien, H. L. et Toms, E. G. (2008). What is user engagement ? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6):938–955.

O'Connell, D. C., Kowal, S. et Kaltenbacher, E. (1990). Turn-taking : A critical analysis of the research tradition. *Journal of Psycholinguistic Research*, 19(6):345–373.

Oertel, C., Looze, C., Scherer, S., Windmann, A., Wagner, P. et Campbell, N. (2011). Towards the Automatic Detection of Involvement in Conversation. *In* Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C. et Nijholt, A., éditeurs : *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*, volume 6800 de *Lecture Notes in Computer Science*, pages 163–170. Springer Berlin Heidelberg.

Ohshima, N., Kimijima, K., Yamato, J. et Mukawa, N. (2015). A conversational robot with vocal and bodily fillers for recovering from awkward silence at turn-takings. *In 2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 325–330.

Ortony, A. (2003). On making believable emotional agents believable. *In* Trappl, R., éditeur : *Emotions in Humans and Artifacts*, pages 189–212.

Peters, C., Pelachaud, C., Bevacqua, E., Mancini, M. et Poggi, I. (2005). A Model of Attention and Interest Using Gaze Behavior. *In* Panayiotopoulos, T., Gratch, J., Aylett, R., Ballin, D., Olivier, P. et Rist, T., éditeurs : *Proceedings of 5th International Working Conference, IVA 2005*, volume 3661 de *Lecture Notes in Computer Science*, pages 229–240, Kos, Greece. Springer Berlin Heidelberg.

Pfeifer, R. et Pitti, A. (2012). *La révolution de l'intelligence du corps*. Paris : Manuella editions édition.

Picard, R. W. (2000). *Affective Computing*. Cambridge, MA : The MIT Press, reprint édition.

Pimentel, M. G., Fuks, H. et Lucena, d. C. (2004). Mediated Chat 2.0 : Embedding Coordination into Chat Tools. *In Proceedings of COOP*, volume 4, pages 99–103.

Pinchbeck, D. (2008). *Trigens can't swim : intelligence and intentionality in first person game worlds*, pages 242–260. Potsdam University Press, Potsdam.

Pylyshyn, Z. W. (1981). Computation and cognition : Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, 3(01):111–132.

Rabin, M. (1991). Incorporating Fairness Into Game Theory. *The American Economic Review*, 83(5).

Rao, A. S. et Georgeff, M. P. (1991). Modeling Rational Agents within a BDI-Architecture. *In* Allen, J., Fikes, R. et Sandewall, E., éditeurs : *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR-91)*, pages 473–484, San Francisco, CA.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review*, 85(2):59–109.

Ratcliff, R. (1980). A note on modeling accumulation of information when the rate of accumulation changes over time. *Journal of Mathematical Psychology*, 21(2):178–184.

Ratcliff, R. et McKoon, G. (2008). The Diffusion Decision Model : Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, 20:873–922.

Ratcliff, R. et Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology : Human perception and performance*, 26(1):127.

Raux, A. et Eskenazi, M. (2012). Optimizing the Turn-taking Behavior of Task-oriented Spoken Dialog Systems. *ACM Transaction on Speech Language Processing*, 9(1):1 :1–1 :23.

Ravenet, B., Cafaro, A., Biancardi, B., Ochs, M. et Pelachaud, C. (2015). Conversational Behavior Reflecting Interpersonal Attitudes in Small Group Interactions. *In Intelligent Virtual Agents : 15th International Conference, IVA 2015, Delft, The Netherlands, August 26-28, 2015, Proceedings*, volume 9238, page 375. Springer.

Reeves, B. et Nass, C. (1996). *The Media Equation : How People Treat Computers, Television, and New Media Like Real People and Places*. New York, NY, USA : Cambridge University Press.

Reidsma, D., Kok, I. d., Neiberg, D., Pammi, S. C., Straalen, B. v., Truong, K. et Welbergen, H. v. (2011). Continuous interaction with a virtual human. *Journal on Multimodal User Interfaces*, 4(2):97–118.

Rich, C., Ponsler, B., Holroyd, A. et Sidner, C. L. (2010). Recognizing Engagement in Human-Robot Interaction. *In Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Osaka,Japan.

Richardson, M. J., Shockley, K., Fajen, B. R., Riley, M. A. et Turvey, M. T. (2008). Six Principles for an Embodied–Embedded Approach to Behavior. *In* Calvo, P. et Gomil, T., éditeurs : *Handbook of Cognitive Science : An Embodied Approach*, pages 161–187. Elsevier.

Rickel, J. et Johnson, W. L. (1997). Steve : An animated pedagogical agent for procedural training in virtual environments. *Intelligent virtual agents, Proceedings of Animated Interface Agents : Making Them Intelligent*, pages 71–76.

Riedl, M. O. et Young, R. M. (2005). An Objective Character Believability Evaluation Procedure for Multi-agent Story Generation Systems. *In* Panayiotopoulos, T., Gratch, J., Aylett, R., Ballin, D., Olivier, P. et Rist, T., éditeurs : *Intelligent Virtual Agents*, numéro 3661 de Lecture Notes in Computer Science, pages 278–291. Springer Berlin Heidelberg. DOI : 10.1007/11550617_24.

Riest, C., Jorschick, A. B. et de Ruiter, J. P. (2015). Anticipation in turn-taking : mechanisms and information sources. *Language Sciences*, 6:89.

Rio, K. W., Rhea, C. K. et Warren, W. H. (2014). Follow the leader : Visual control of speed in pedestrian following. *Journal of Vision*, 14(2):4–4.

Rozendaal, M., Keyson, D., Ridder, H. et Craig, P. (2009). Game feature and expertise effects on experienced richness, control and engagement in game play. *AI and Society*, 24(2):123–133.

Sacks, H., Schegloff, E. A. et Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, pages 696–735.

Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in society*, 29(01):1–63.

Scherer, S., Marsella, S. C., Stratou, G., Xu, Y., Morbini, F., Egan, A., Rizzo, A. et Morency, L.-P. (2012). Perception Markup Language : Towards a Standardized Representation of Perceived Nonverbal Behaviors. *In The 12th International Conference on Intelligent Virtual Agents (IVA)*. Santa Cruz, CA.

Schlangen, D., Baumann, T., Buschmeier, H., Buss, O., Kopp, S., Skantze, G. et Yaghoubzadeh, R. (2010). Middleware for incremental processing in conversational agents. *In Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 51–54. Association for Computational Linguistics.

Schlangen, D. et Skantze, G. (2011). A General, Abstract Model of Incremental Dialogue Processing. *Dialogue & Discourse*, 2(1):83–111.

Schmidt, R. A. (1975). A schema theory of discrete motor skill learning. *Psychological Review*, 82(4):225–260.

Sebanz, N., Bekkering, H. et Knoblich, G. (2006). Joint action : bodies and minds moving together. *Trends in Cognitive Sciences*, 10(2):70–76.

Selfridge, E., Arizmendi, I., Heeman, P. et Williams, J. (2013). Continuously predicting and processing barge-in during a live spoken dialogue task. *In Proceedings of the SIGDIAL 2013 Conference*, pages 384–393.

Sellen, A. J. (1995). Remote Conversations : The Effects of Mediating Talk with Technology. *Human-Computer Interaction*, 10(4):401–44.

Short, J., Williams, E. et Christie, B. (1976). *The Social Psychology of Telecommunications*. London : Wiley.

Sidner, C. L., Lee, C., Lesh, N. et Kidd, C. D. (2004). Where to Look : A Study of Human-Robot Interaction. *In Proceedings of Intelligent User Interfaces*, pages 78–84, Madeira, Portugal.

Skantze, G. et Hjalmarsson, A. (2010). Towards incremental speech generation in dialogue systems. *In Proceedings of SIGDIAL 2010*, pages 1–8. Association for Computational Linguistics.

Skantze, G., Hjalmarsson, A. et Oertel, C. (2014). Turn-taking, feedback and joint attention in situated human–robot interaction. *Speech Communication*, 65:50–66.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., Ruiter, J. P. d., Yoon, K.-E. et Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

Ter Maat, M., Truong, K. P. et Heylen, D. (2010). How turn-taking strategies influence users' impressions of an agent. *In Intelligent Virtual Agents*, pages 441–453. Springer.

Thórisson, K. R. (1999). A Mind Model for Multimodal Communicative Creatures & Humanoids. *International Journal of Applied Artificial Intelligence*, 13(4):449–486.

Thórisson, K. R. (2002). Natural turn-taking needs no manual : Computational theory and model, from perception to action. *Multimodality in language and speech systems*, 19.

Thórisson, K. R., Gislason, O., Jonsdottir, G. R. et Thorisson, H. T. (2010). A multiparty multimodal architecture for realtime turntaking. *In Intelligent Virtual Agents*, pages 350–356. Springer.

Torreira, F., Bögels, S. et Levinson, S. C. (2015). Breathing for answering : the time course of response planning in conversation. *Frontiers in Psychology*, 6(284).

Tuomela, R. (1993). What is cooperation ? *Erkenntnis*, 38(1):87–101.

van Vugt, H. C., Konijn, E. A., Hoorn, J. F., Keur, I. et Eliëns, A. (2007). Realism is not all ! User engagement with task-related interface characters. *Interacting with Computers*, 19:267 – 280.

Von Holst, E. (1973). *The Behavioural Physiology of Animals and Man : The Collected Papers of Erich Von Holst*. Miami, FL : University of Miami Press.

von Neumann, J. et Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton University Press.

Vygotsky, L. (1978). *Mind in Society - Development of Higher Psychological Processes*. Harvard University Press, Cambridge, new ed édition.

Ward, N. et Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32(8):1177–1207.

Ward, N. G., Rivera, A. G., Ward, K. et Novick, D. G. (2005). Root causes of lost time and user stress in a simple dialog system. *In Proceedings of INTERSPEECH 2005*.

Warren, W. H. (2006). The Dynamics of Perception and Action. *Psychological Review*, 113(2):358–389.

Wilson, M. et Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review*, 12(6):957–968.

Wilson, T. P. et Zimmerman, D. H. (1986). The structure of silence between turns in two party conversation. *Discourse Processes*, 9(4):375–390.

Witt, S. (2014). Modeling user response timings in spoken dialog systems. *International Journal of Speech Technology*, 18(2):231–243.

Yngve, V. H. (1970). On getting a word in edgewise. *Chicago Linguistics Society, 6th Meeting*, pages 567–578.

Zwiers, J., Welbergen, H. v. et Reidsma, D. (2011). Continuous Interaction within the SAIBA Framework. *In* Vilhjálmsson, H. H., Kopp, S., Marsella, S. et Thórisson, K. R., éditeurs : *Intelligent Virtual Agents*, numéro 6895 de Lecture Notes in Computer Science, pages 324–330. Springer Berlin Heidelberg.