Architecture d'agent conversationnel animé pour la coordination émergente des tours de parole avec un utilisateur

Mathieu Jégou[1], Pierre Chevaillier[1,2]

1. IRT b-com,
25 rue Claude Chappe, F-29280 Plouzané, France
mathieu.jegou@b-com.com

2. ENIB, Lab-STICC,
25 rue Claude Chappe, F-29280 Plouzané, France
pierre.chevaillier@enib.fr

RÉSUMÉ. Cet article présente un modèle continu et émergent pour la coordination des tours de parole de dyades agent-utilisateur. La particularité de ce modèle réside dans la capacité de l'agent à gérer de manière incrémentale un grand nombre de situations observées dans les interactions humaines que ce soit des transitions fluides ou des moments de conflit. Nous commençons par présenter notre modèle conceptuel puis nous introduisons l'implémentation de notre modèle. Nous évaluons enfin notre modèle par une expérimentation en magicien d'Oz où des utilisateurs dialoguent en temps réel avec l'agent. Les résultats de cette évaluation ont montré la capacité de notre modèle à améliorer la coordination de l'agent avec l'utilisateur. Néanmoins, nous avons aussi observé une difficulté des utilisateurs à lier le comportement de l'agent à ses intentions vis-à-vis de la coordination des tours de parole.

ABSTRACT. This article presents a continuous and emergent model for the speaking turns coordination of user-agent dyads. The peculiarity of this model resides in its ability to manage incrementally a large number of situations observed in human spoken interactions, competitive overlaps or smooth transitions. We begin this paper by introducing our conceptual model then we introduce the implementation of our model. We finally evaluate our model by a wizard-of-oz experiment where users interact in real-time with the agent. The results of this evaluation shows that our model improves the agent's ability to coordinate its speaking turns with the user. However, we also observe that users have difficulties to link the agent's behavior to its intentions towards turn-taking management.

MOTS-CLÉS : agent conversationnel, architecture comportementale, coordination, perception-action, tour de parole, conflits de parole, prosodie.

KEYWORDS: conversational agent, behavioral architecture, coordination, perception-action, turn-taking, prosody.

DOI:10.3166/RIA.31.541-608 © 2017 Lavoisier

Revue d'intelligence artificielle – n° 5/2017, 581-608

1. Introduction

Selon Cassell *et al.* (2000), les agents conversationnels animés sont des entités graphiques 2D ou 3D avec une apparence anthropomorphique, capables de dialoguer de manière naturelle avec l'utilisateur en reconnaissant et produisant des signaux verbaux et non-verbaux. Afin d'assurer une interaction naturelle avec l'utilisateur, ces agents conversationnels animés doivent disposer de plusieurs capacités dont la capacité à coordonner leurs tours de parole avec l'utilisateur. La coordination des tours de parole se réfère à la capacité des participants à alterner les moments de parole et d'écoute de sorte que, la majorité du temps, un seul participant a la parole (Sacks *et al.*, 1974).

Dans un grand nombre de ces systèmes, les échanges de tour de parole entre utilisateurs et agents sont rigides, loin de la fluidité observée dans les interactions humaines. En effet, pour une interaction d'initiative mixte spontanée, efficace et engageante avec l'utilisateur, le module de gestion du tour de parole doit être capable de gérer la prise de parole de manière optimale, mais aussi de gérer les interruptions, les recouvrements non compétitifs ou encore les silences longs. À notre connaissance, chaque problématique citée ci-dessus a été traitée de manière séparée par des auteurs différents. Jonsdottir et Thórisson (2013), par exemple, ont élaboré un agent capable d'apprendre à détecter la fin de tour de l'utilisateur au cours de l'interaction avec ce dernier ; Selfridge *et al.* (2013), ont proposé un algorithme permettant de différencier des tentatives d'interruption de la part de l'utilisateur d'énoncés prononcés par ce dernier non dirigés vers le système ; Ohshima *et al.* (2015) ont créé un agent utilisant des stratégies pour relancer l'interaction avec l'utilisateur en cas de silences longs avec ce dernier. Aucun auteur n'a intégré l'ensemble de ces problématiques en un modèle de contrôle du comportement d'un agent. Une autre problématique réside dans la prise en compte de facteurs liés au dialogue, à la personnalité et aux attitudes interpersonnelles des participants dans la modulation de leur comportement vis-à-vis de la coordination des tours de parole. Un agent capable de moduler la manière dont il prend, garde ou laisse la parole à l'utilisateur selon ces facteurs, améliorerait grandement le caractère naturel de l'interaction avec l'utilisateur.

Sur la base d'un modèle élaboré et précédemment présenté (Jégou *et al.*, 2015), nous présentons dans cet article un module de coordination des tours de parole d'un agent conversationnel en interaction temps réel avec un utilisateur, montrant la capacité d'un agent à coordonner sa parole avec un utilisateur tout en modulant la manière dont il prend, garde ou laisse la parole selon une variable que nous appelons de manière volontairement abstraite « motivation à changer de rôle ».

Cet article se structure comme suit. Nous présentons dans une première partie (section 2) un état de l'art des travaux récents sur le contrôle du tour de parole utilisateur-agent, à partir duquel nous justifions notre positionnement. La section 3 présente le modèle conceptuel de tour de parole que nous avons développé. La section 4 décrit la manière dont nous avons implémenté notre modèle de coordination des tours de parole dans une architecture d'agent inspirée d'Ymir (Thórisson, 1999) et d'ASAP (Kopp *et al.*, 2014), et montre son fonctionnement

dans le cadre d'un scénario d'interaction temps-réel entre un utilisateur et un agent. La section 5 présente une étude expérimentale, où, dans le cadre d'un scénario de négociation entre un agent et un utilisateur, nous avons comparé notre modèle à un algorithme simple de gestion des tours utilisant des temporisations.

2. Positionnement

2.1. Tour de parole dans les interactions humaines

2.1.1. Définition de la coordination des tours de parole

Dans une conversation dyadique, les participants humains alternent leurs tours de parole afin que, la majorité du temps, un seul participant ait la parole (Sacks *et al.,* 1974). L'intervalle où un seul participant a la parole est appelé un tour de parole, le possesseur d'un tour est appelé le locuteur, l'autre participant étant appelé l'auditeur (Sacks *et al.,* 1974). Le processus par lequel les participants garantissent cette alternance de tour est appelé coordination des tours de parole (*turn-taking* en anglais (Sacks *et al.,* 1974)).

Selon (Sacks *et al.,* 1974), cette coordination des tours de parole est gérée localement, les possesseurs des tours n'étant pas attribués à l'avance, et par l'interaction entre les participants qui, chacun, peuvent produire plusieurs types d'actions observables, dans des conversations dyadiques, selon Bunt et Girard (2005) :

Prendre le tour : l'auditeur courant décide de devenir le possesseur du tour alors que ce dernier est disponible.

Saisir le tour : l'auditeur courant décide de devenir le locuteur courant alors que ce dernier n'a pas fini son tour.

Garder le tour : le locuteur courant souhaite rester le possesseur du tour.

Laisser le tour : le locuteur courant finit son tour et laisse un autre participant prendre le tour à sa suite.

Une des propriétés remarquables de la coordination des tours de parole réside dans la fluidité avec laquelle les participants s'échangent la parole. En effet, les transitions de tour, c'est-à-dire, les moments où le possesseur de tour change, s'effectuent de manière fluide, ne durant en général que quelques millisecondes (Heldner et Edlund, 2010). De même, les participants résolvent rapidement (1 à 2 secondes) les moments de conflits entre les participants (Sacks *et al.,* 1974).

Plusieurs approches ont tenté d'expliquer cette fluidité dans la coordination des tours de parole. Sacks *et al.* (1974) considèrent que cette coordination fine des tours de parole est due à la capacité des auditeurs à percevoir, dans le discours du locuteur courant des éléments syntaxiques ou prosodique leur permettant de prédire le moment où une transition peut potentiellement avoir lieu. Néanmoins, peu d'études ont mis permis d'identifier les différents éléments permettant aux participants de projeter la fin de tour du locuteur courant.

Contrairement à Sacks *et al.* (1974), Duncan (1972) postule que la coordination des tours de parole s'effectue principalement par l'échange d'un certain nombre de signaux verbaux et non-verbaux par lesquels les participants informent leurs interlocuteurs de leurs actions (prendre le tour, saisir le tour, garder le tour, libérer le tour) envers le tour de parole. Duncan (1972) a initialement identifié plusieurs types de signaux permettant d'informer les autres participants de leurs actions envers le tour de parole.

Plusieurs études ont, depuis, complété cet ensemble de signaux initial, par d'autres signaux verbaux et non-verbaux. Les variations de regard sont par exemple utilisées par le locuteur courant regardant vers l'auditeur pour lui signifier sa volonté de donner le tour et l'auditeur détournant le regard pour signifier sa prise de tour (Oertel *et al.*, 2013).

Dans cet article, nous avons choisi de nous focaliser sur une coordination des tours de parole sur la base des signaux prosodiques, c'est-à-dire une augmentation d'énergie acoustique et de hauteur de voix utilisé par le locuteur et l'auditeur respectivement pour garder et saisir le tour (Kurtić *et al.*, 2013) et une baisse d'énergie acoustique et de hauteur de voix, utilisé par le locuteur courant pour laisser le tour, ces signaux de coordination faisant partie des signaux les plus étudiés par les auteurs (Gravano et Hirschberg, 2011) . De plus, l'emploi de ces signaux co-verbaux en plus du contenu des énoncés prononcés par les participants semble suffire aux participants pour avoir une coordination des tours de parole (Sellen, 1995), bien que celle-ci diffère de ce qui peut être observé lorsque les participants ont accès aux signaux non-verbaux de leur partenaires .

Selon Duncan (1972) et Sacks *et al.* (1974), la coordination des tours de paroles sert avant tout à garantir l'alternance des tours de sorte d'éviter les silences trop longs et les paroles simultanées conflictuelles entre les participants. Plusieurs auteurs ont critiqué cette vision du tour de parole comme un ensemble de normes à suivre pour garantir le succès d'une conversation (voir O'Connell *et al.* (1990)).

En effet, plusieurs études ont montré que certaines situations liées à la coordination des tours de parole, telles que les interruptions (Goldberg, 1990), ou les moments de silences intra-tour longs, c'est-à-dire d'une seconde ou plus (Roberts et Francis, 2013) ont une fonction précise dans la conversation. Ainsi les interruptions, c'est-à-dire des situations où l'auditeur courant saisit le tour alors que le locuteur n'avait pas fini son tour (Goldberg, 1990), peuvent être compétitives, servant à montrer une attitude dominante envers le locuteur courant (Ter Maat *et al.*, 2010; Cafaro *et al.*, 2016), ou coopératives, servant à montrer une attitude amicale envers le locuteur courant (Goldberg, 1990). De la même manière, Roberts et Francis (2013) montrent que le futur locuteur laisse souvent un silence plus long lorsqu'il s'apprête à prendre le tour pour montrer son désaccord avec le locuteur courant (Roberts et Francis, 2013).

Enfin, une dernière approche considère qu'une part de la coordination des tours de parole est liée à un phénomène de couplage dans la production des signaux entre les participants, le couplage étant ici défini comme une influence mutuelle continue entre deux participants dans la production de leur actions (Bevacqua *et al.*, 2014).

Plusieurs auteurs ont ainsi montré des effets d'entrainement dans le comportement des participants engagés dans les conversations. Levitan *et al.* (2015) observent par exemple un alignement dans les durées de silence laissées par les participants avant de prendre le tour. D'autres auteurs ont observé des synchronisations dans les cycles de respirations du locuteur courant et du futur locuteur à l'approche de la fin de tour du locuteur courant (McFarland, 2001**)**. Wilson et Wilson (2005) proposent un modèle où la coordination des tours de parole entre les participants est expliquée par une synchronisation d'oscillateurs cérébraux endogènes liés au cycle de prononciation des syllabes entre le locuteur courant et le locuteur suivant. Ils se basent sur le fait que les durées de silence inter-tour tendent à être un multiple d'une unité de temps tel qu'observé par McFarland (2001) en anglais et Bailly et Gouvernayre (2012) en français.

Ces études tendent à montrer que les participants adaptent en continu leurs propres actions et productions de signaux au comportement de leur partenaire. Selon McFarland (2001), les productions de signaux et actions des participants envers la coordination des tours de parole seraient partiellement résultantes d'une influence directe des signaux produits par l'interlocuteur, et les actions des participants seraient alors une propriété émergente, c'est-à-dire provenant de l'interaction entre les participants plus que des prises de décision individuelles des participants.

2.1.2. Modèles informatiques de coordination des tours de parole

La majorité des modèles de gestion du tour de parole pour des dyades humain-agent utilisent des règles simples pour la coordination des participants: ne jamais parler en même temps que l'utilisateur, se taire lorsque l'utilisateur interrompt le système, et prendre le tour le plus rapidement possible après la fin de tour de l'utilisateur (Jonsdottir et Thórisson, 2013 ; Selfridge *et al.*, 2013). Selon ces approches, la fluidité de l'interaction repose sur la capacité de l'agent à prendre le tour le plus rapidement possible après la fin de tour de l'utilisateur, tout en évitant les chevauchements accidentels liés à une mauvaise estimation de la fin de tour de l'utilisateur ou à une pause dans le discours de ce dernier interprétée comme une fin de tour. Plusieurs modèles ont ainsi cherché à optimiser la prise de tour de l'agent afin de reproduire les temps de transitions moyens observés entre les tours de parole dans les conversations humaines (Huang *et al.*, 2011 ; Raux et Eskenazi, 2012 ; Jonsdottir et Thórisson, 2013). Ces modèles cherchent à minimiser les erreurs de détection de la fin de tour. Ils utilisent de l'apprentissage automatique, soit**,** au préalable sur des corpus d'interactions humaines (De Vault *et al.*, 2011 ; Huang *et al.*, 2011 ; Raux et Eskenazi, 2012) ou en temps réel (Jonsdottir et Thórisson, 2013). Ces modèles s'inspirent majoritairement d'approches de psychologie sociale sur la manière dont des participants humains coordonnent leurs tours dans des conversations. Néanmoins, ces modèles ne tiennent pas compte de facteurs pouvant modifier la manière dont l'agent coordonne ses tours de parole, tels que les attitudes interpersonnelles ou encore les émotions. Tenir compte de ces facteurs dans un modèle de coordination des tours de parole pourrait ainsi améliorer le caractère naturel de l'interaction entre l'agent et l'utilisateur. Certains modèles de coordination des tours de parole pour des interactions utilisateur-agent et agent-

agent ont ainsi introduit des variables liées à l'importance (Lessmann *et al.*, 2004 ; Selfridge et Heeman, 2009 ; Thórisson *et al.*, 2010), à la nature de l'énoncé que doit prononcer l'agent (Cafaro *et al.*, 2016) ou encore liées aux attitudes interpersonnelles des participants (Ravenet *et al.*, 2015). Ainsi selon le degré d'importance de son énoncé, de la nature de sa contribution ou de celle de son partenaire, et de son attitude, soumise ou dominante, envers ce dernier, l'agent garde ou non la parole si un autre agent ou utilisateur cherche à l'interrompre et de la même manière interrompt ou non le locuteur courant.

2.1.3. Positionnement

Les modèles actuels conçoivent la gestion du tour de parole comme un ensemble d'actions, provenant d'intentions communicatives explicitement formulées par l'agent et donnant lieu au comportement final de l'agent. Dans les approches traditionnelles, l'agent évalue le comportement de son partenaire puis, lorsqu'il détecte la fin de tour de ce dernier, décide de prendre le tour. Une fois décidé de prendre le tour ce dernier prend effectivement le tour. Des variantes peuvent être mises en places, telles que proposés par Ravenet *et al.* (2015) : l'agent peut choisir d'attendre la fin de tour ou d'interrompre ce dernier, mais une fois la décision prise l'agent le comportement est effectivement réalisé par l'agent.

Cela exclut le fait que la gestion des tours de parole est un processus nécessitant une adaptation continue du comportement de chaque agent au comportement de l'autre pour qu'une coordination effective ait lieu entre les participants (McFarland, 2001 ; Thórisson, 2002). Ainsi, un participant peut décider de laisser la parole puis quelques millisecondes plus tard de se raviser et continuer à parler, dans un tel cas l'autre participant ayant décidé de prendre le tour à la vue de la fin de tour de l'utilisateur devra rapidement réagir à la décision de reprendre la parole du locuteur précédent (Thórisson, 2002). De même, le comportement final des participants en situation de recouvrement compétitif n'est pas décidé à l'avance mais résolu au cours du recouvrement. Par exemple, l'auditeur précédent, ayant cherché à interrompre le locuteur courant, peut décider à mesure que le conflit dure ou que le locuteur insiste de plus en plus pour garder la parole (augmente sa voix) de se raviser. De même, le locuteur courant, percevant la tentative d'interruption de l'auditeur courant peut décider d'augmenter son énergie acoustique ou sa hauteur de voix pour tenter d'empêcher l'auditeur courant de prendre la parole, puis à mesure que le conflit dure ou que l'auditeur insiste de plus en plus pour prendre la parole peut décider de se raviser. Ainsi de notre point de vue, le comportement final d'un agent ne peut être déterminé à l'avance, planifié par l'agent, mais un résultat de l'interaction avec son partenaire.

L'agent a une motivation initiale de prendre, laisser ou garder le tour, dépendant de différents facteurs comme le contenu des énoncés échangés par les participants, les attitudes interpersonnelles ou les émotions. Cette motivation initiale fixe un but que l'agent cherche à atteindre en modifiant sa propre production de signaux pour atteindre ce but. Néanmoins, l'agent adapte en continu son action et sa production de signaux au comportement perçu de son partenaire, il en résulte que le comportement final de l'agent, est le fruit d'une interaction complexe entre les buts de l'agent

(prendre, laisser, garder le tour) et l'adaptation de son comportement au comportement de son partenaire.

Cette vision du tour de parole où l'agent a un but initial de changer ou non de rôle mais dont la réalisation concrète du comportement de l'agent est le fruit de l'interaction entre l'agent et son interlocuteur se rapproche de modèles de psychologie cognitive tels que le modèle de dynamique comportementale énoncé par Warren (2006). De manière similaire à notre formulation du problème de la gestion du tour de parole, Warren (2006) considère que le contrôle du comportement d'un agent biologique se fait à deux niveaux. Au premier niveau, l'agent fixe ses buts comportementaux tels que marcher vers une cible. Au second niveau, l'agent est engagé dans un cycle de perception-action avec son environnement, le comportement de l'agent n'est pas planifié à l'avance mais est le résultat d'une adaptation continue des actions de l'agent en fonction des informations reçues de l'environnement. Pour cette raison, nous avons choisi de nous baser sur le cadre théorique de la dynamique comportementale pour élaborer notre modèle.

3. Modèle théorique

Avant de décrire dans les détails notre approche, nous souhaitons tout d'abord préciser la distinction que nous faisons entre les différents processus cognitifs impliqués dans la coordination des tours de parole.

La figure 1 présente les deux différentes couches impliquées dans le contrôle du comportement de l'agent. La première concerne les facteurs contextuels qui impactent la manière dont l'agent prend, laisse ou garde le tour. Ces facteurs contextuels sont nombreux, l'importance de la contribution de l'agent envers la progression du dialogue (Selfridge et Heeman, 2009 ; Thórisson *et al.*, 2010), la nature de cette contribution, coopérative ou compétitive (Cafaro *et al.*, 2016), ou encore l'attitude de l'agent envers son partenaire (Ter Maat *et al.*, 2010 ; Ravenet *et al.*, 2015) sont des exemples de facteurs contextuels. Dans notre modèle, ces facteurs n'impactent pas directement le comportement de l'agent en matière de stratégies de prise de parole mais modulent la motivation de l'agent à changer de rôle (passer de locuteur à auditeur ou inversement). Ce que nous appelons motivation, ici, est similaire à ce que d'autres auteurs appellent « urgence à parler » (Thórisson *et al.*, 2010) ou « intention à parler » (Lessmann *et al.*, 2004) dans le sens où cette variable module la force avec laquelle l'agent cherche à prendre ou laisser la parole. Nous avons choisi d'appeler cette variable *motivation* pour rester le plus neutre possible envers les différents facteurs contextuels qui pourraient moduler cette variable. Ainsi, prendre le tour tôt n'est pas toujours lié à une urgence à parler et n'est pas toujours lié à des intentions communicatives explicites, les émotions et les attitudes du participant pouvant aussi impacter le comportement de l'agent. Dans notre modèle, la motivation à changer de rôle est une variable continue qui varie entre -1 et 1, -1 voulant dire que l'agent a une motivation forte à garder son rôle et 1 voulant dire que l'agent a une motivation forte à changer de rôle. Entre ces deux valeurs, l'agent peut continuellement varier la force avec laquelle il cherchera à prendre ou garder son rôle. Plus la motivation est proche de 0, moins la force avec

laquelle le participant cherchera à changer ou garder son rôle courant sera grande. Dans de telles situations, l'agent abandonne plus rapidement ses tentatives de changement de rôle et il prend plus en compte le comportement de son partenaire, tel qu'il peut le percevoir.

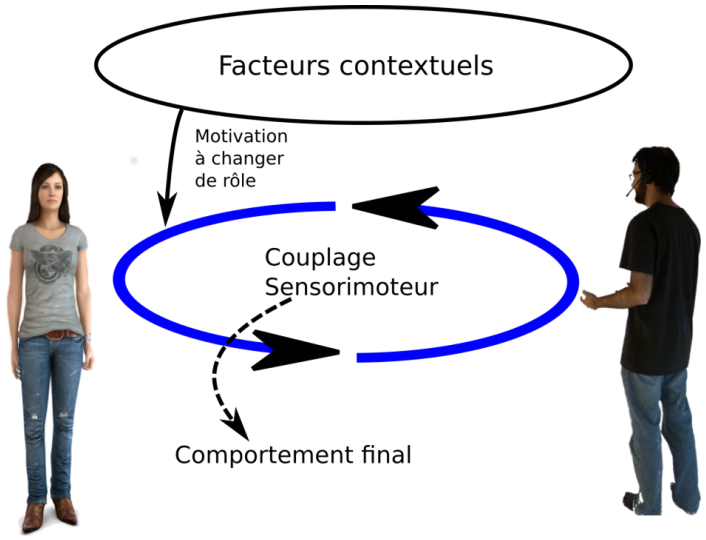

Figure 1. Illustration de la distinction entre les deux couches de contrôle du comportement de l'agent impliquées dans la coordination des tours de parole

La valeur de motivation ne détermine pas à elle seule le comportement final de l'agent en matière de prise, fin ou continuation de tour. Cette variable représente davantage les buts comportementaux de l'agent (prendre, laisser, garder le tour), alors que le comportement de l'agent est une propriété émergente de l'interaction entre lui et son partenaire. Elle résulte d'un ajustement continu des productions de l'agent qui opère à un second niveau de l'interaction, le niveau sensorimoteur (figure 1). Ce niveau rend compte du contrôle exercé par l'agent sur les signaux qu'il produit en regard de sa motivation initiale (par exemple détourner le regard s'il souhaite prendre la parole). L'agent étant continuellement couplé avec son partenaire, la production de signaux de son interlocuteur influence directement la manière dont l'agent va varier ses propres signaux, et inversement, la variation de ses actions influencera le comportement de son partenaire. Il en résulte que le comportement final de l'agent, c'est-à-dire si le participant va prendre, laisser ou garder le tour, et la forme du comportement (les signaux produits par l'agent) sont des propriétés émergentes de l'interaction complexe et continue entre la motivation de l'agent et la perception des signaux de son interlocuteur.

Pour cet article, nous supposons que la valeur de motivation est influencée par différentes variables contextuelles, mais, dans cet article, ne nous intéressons pas

plus précisément à la manière dont celle-ci varie selon ces différentes variables. L'étude de la manière dont les facteurs contextuels influent sur la motivation à changer de rôle est une perspective de notre travail. Nous nous intéressons ainsi uniquement à la manière dont l'agent varie ses signaux à partir de sa motivation et de la perception des signaux produits par son partenaire. De même, dans cette section, afin de rester générique, nous considérons que les participants produisent et perçoivent des signaux théoriques provenant de leur partenaire, sans nous intéresser pour le moment à la nature de ces signaux. Nous verrons dans la section 4, comment nous avons appliqué notre modèle à la perception et à la production des signaux co-verbaux d'énergie acoustique et de hauteur de voix.

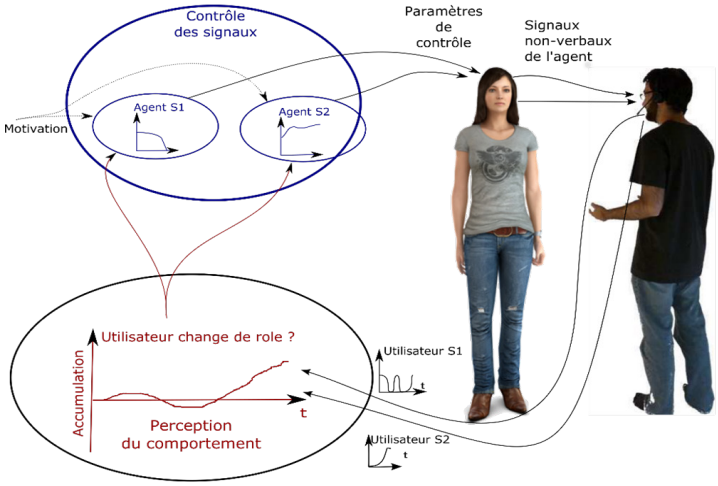

Figure 2. Schéma du modèle, avec les deux composantes : la composante de prise de décision perceptive et la composante de contrôle des signaux

La figure 2 illustre l'organisation générale de notre modèle qui est structuré en deux composantes : perception du comportement et contrôle (de la production) des signaux. Nous l'illustrons sur cette figure en prenant l'exemple d'un utilisateur produisant deux types de signaux non-verbaux que l'agent peut percevoir, US1 et US2, et d'un agent produisant lui aussi deux types de signaux AS1 et AS2.

La composante de perception du comportement a été réalisée selon les principes du modèle de dérive-diffusion de Ratcliff (1978), un modèle de psychologie cognitive utilisé pour modéliser la prise de décision perceptive d'agents devant discriminer entre deux alternatives la nature d'une information provenant de leur

environnement. Cette composante agrège les différents signaux produits par son partenaire et met à jour continuellement une variable d'accumulation renseignant son niveau de certitude au sujet du comportement de son partenaire, c'est-à-dire si l'agent perçoit que son partenaire est en train de changer de rôle ou, au contraire, qu'il est en train de garder son rôle. Cette valeur d'accumulation varie entre -1 et 1. -1 signifie que l'agent a une certitude forte envers le fait que son partenaire est en train de garder son rôle et 1 qu'il a une certitude forte envers le fait que son partenaire est en train de changer de rôle. Lorsque cette certitude atteint la valeur 1 l'agent décide de changer de rôle. L'équation (1) est la formulation générale du module de perception du comportement de l'agent :

$$\frac{\mathbf{d\gamma(t)}}{\mathbf{dt}} = \alpha(\mathrm{t}) + \mathrm{c} \times \frac{\mathrm{dW}}{\mathrm{dt}} \quad (1)$$

Dans l'équation (1), $\gamma(\mathrm{t})$ est la valeur d'accumulation, représentant un niveau de certitude concernant le comportement du partenaire (passer de locuteur à auditeur ou inversement), $\alpha(\mathrm{t})$ le taux d'accumulation variant selon les signaux du partenaire, tels que l'agent les perçoit et un terme stochastique variant selon une loi normale centrée de variance c. $\frac{\mathrm{dW}}{\mathrm{dt}}$ représente un bruit aléatoire suivant une loi normale de moyenne 0 et de variance 1.

Le taux d'accumulation est fonction des variations de signaux produits par le partenaire de l'agent. Cette fonction a une forme générale définie par l'équation :

$$\alpha(\mathrm{t}) = \Sigma_{j=0}^{n_s} a_j(\dot{s_j}(t), s_j(t)) \quad (2)$$

Dans l'équation (2), $\alpha(t)$ varie de manière continue selon une somme de processus d'accumulation partiels, qui calcule un taux d'accumulation pour chaque signal que l'agent perçoit, $\alpha > 0$ indique que les variations de signaux produits par le partenaire sont en faveur de l'hypothèse que le partenaire est en train de changer de rôle et $\alpha < 0$ indique que les variations de signaux perçus sont en faveur de l'hypothèse d'un partenaire gardant son rôle courant, n_s représente le nombre de signaux interprétés par l'agent, $s_j(t)$ représente la valeur à l'instant t du signal j (énergie acoustique ou pitch par exemple).

La seconde composante de notre modèle contrôle la production des signaux de l'agent. Elle définit comment les signaux de l'agent varient selon la valeur d'accumulation γ calculée dans la composante de perception du comportement du partenaire et la propre motivation à changer de rôle de l'agent. Le module de contrôle des signaux non-verbaux de l'agent a été implémenté selon les principes de la dynamique comportementale. Cela implique que, d'après la dynamique comportementale, la production d'action de l'agent est influencée par deux variables, d'une part ses buts, représenté par la motivation m à changer de rôle, et d'autre part directement par l'information provenant de l'environnement, ici représenté par la valeur d'accumulation γ de l'agent envers le comportement de son partenaire. Ici, l'agent contrôle chacun de ses signaux selon une équation différentielle. La motivation m et la valeur d'accumulation γ vont contrôler

l'attracteur de ces équations, c'est-à-dire la valeur finale vers laquelle converge chaque signal produit par l'agent.

Ainsi, par exemple, l'équation de contrôle de l'énergie acoustique de l'agent rend compte de la manière dont l'agent augmente de plus en plus son énergie acoustique à mesure que sa valeur d'accumulation devient de plus en plus forte vis-à-vis du fait que son partenaire est en train de changer de rôle, et qu'il est motivé à garder son tour, ce que l'on observe, par exemple, lorsque le locuteur courant veut garder la parole alors que l'auditeur courant tente de la prendre. L'équation (3) est la formulation générale des équations de contrôle des signaux de l'agent :

$$\frac{\mathrm{d}^2\mathrm{a_j(t)}}{\mathrm{dt}} = -\mathrm{b} \times \frac{\mathrm{da_j(t)}}{\mathrm{dt}} - \mathrm{k_g} \times (\mathrm{a_j(t)} - \mathrm{f(m(t), \gamma(t))}) \quad (3)$$

$a_j(t)$ représente la valeur du signal produit, $\frac{\mathrm{d}^2\mathrm{a_j}(t)}{\mathrm{dt}}$ la dérivée seconde du signal, - b un paramètre d'inertie à produire les signaux, $\mathrm{k_g} \times (\mathrm{a_j(t)} - \mathrm{f(m(t), \gamma(t))})$ le terme modulant l'attracteur de l'équation c'est-à-dire la valeur finale vers laquelle l'équation va converger, $m(t)$ la motivation de l'agent à un instant t, $\gamma(t)$ la valeur d'accumulation de l'agent à un instant t et $f(m(t), \gamma(t))$ une fonction définissant, pour chaque signal, selon la motivation et la valeur d'accumulation, l'attracteur de l'équation.

m(t) la motivation à changer de rôle dépend de nombreux facteurs, le contenu des énoncés échangés par les participants (Clark, 1996), les attitudes interpersonnelles des participants (Cafaro *et al.*, 2016) ou encore leurs émotions. Dans cet article nous ne nous préoccupons pas de la manière dont la motivation à changer de rôle varie selon ces différents facteurs, la motivation est donc ici une variable de forçage de notre modèle.

Des simulations d'interactions entre deux agents (Jégou *et al.*, 2015) ont montré la capacité du modèle à faire émerger des scénarios d'interactions tels que des transitions fluides, des moments de conflits et des hésitations à prendre le tour, avec des durées de transition et de conflit proches de ce qu'on observe dans les interactions humaines. Néanmoins, le modèle n'a pas été implémenté dans le cadre d'interactions temps-réel utilisateur-agent. Nous présentons, dans la section suivante, la manière dont nous avons implémenté notre modèle.

4. Conception de l'architecture

Nous présentons, dans cette section, l'implémentation de notre modèle dans une architecture informatique d'agent, inspirée d'ASAP (Kopp *et al.*, 2014) et d'Ymir (Thórisson, 2002) et montrons la capacité de notre modèle à gérer des interactions en temps-réel avec l'utilisateur.

Nous nous appuyons sur des architectures d'agent incrémentales pour l'implémentation de notre modèle (Schlangen *et al.*, 2010 ; Skantze & Hjalmarsson, 2010 ; Kopp *et al.*, 2014). Ces architectures permettent à l'agent de formuler un certain nombre d'hypothèses sur ce qu'est en train de dire l'utilisateur en même

temps que ce dernier est en train de prononcer sa contribution et de réagir directement à ces hypothèses. L'agent peut ainsi avoir une certitude assez forte sur ce que dit l'utilisateur avant que ce dernier ait fini sa phrase, et formuler une réponse avant que l'utilisateur ait fini de parler. L'agent est alors confronté à plusieurs possibilités : commencer à parler alors que l'utilisateur n'a pas fini son tour, c'est-à-dire, l'interrompre, ou attendre que l'utilisateur ait fini de parler pour prendre le tour. Avec notre modèle, le comportement de l'agent est alors guidé par sa composante de gestion du tour de parole qui détermine quand l'agent commencera à réagir à ce que dit l'utilisateur.

Ainsi, telle que définie par notre modèle, la combinaison d'une motivation forte à changer de rôle et d'une perception du comportement de l'utilisateur indiquant que ce dernier insiste peu pour garder son rôle de locuteur fera que l'énoncé de l'agent sera lancé sans attendre la fin de tour de l'utilisateur. D'un autre côté, la combinaison d'une motivation faible à changer de rôle et d'un comportement de l'utilisateur indiquant sa volonté de garder son rôle fera que le lancement de l'énoncé de l'agent ne sera effectué qu'à la fin de l'énoncé de l'utilisateur. Notre modèle a donc vocation à s'appliquer à ce type d'interaction incrémentale. Plusieurs problématiques se posent alors, notamment la capacité à contrôler en continu les différentes productions de signaux non-verbaux générés par notre modèle.

4.1. Principes de l'architecture

La problématique du contrôle continu des signaux de l'agent est partiellement résolue par l'architecture ASAP (Kopp *et al.*, 2014) qui reprend les principes de l'architecture SAIBA (Kopp *et al.*, 2006) pour la génération de l'action avec une architecture séparée en trois modules : le module de Planification d'Intention définit les intentions communicatives de l'agent, tout en fournissant des renseignements sur l'état cognitif de l'agent et sur l'état du dialogue, le module de Planification de Comportement fait correspondre ces intentions communicatives aux actions multimodales générées par l'agent, et le Réaliseur gère l'exécution de l'action. ASAP étend le langage BML de l'architecture SAIBA, utilisé pour envoyer des commandes d'action au Réaliseur. Le langage résultant, le BML*a*, ajoute, entre autres, des capacités à interrompre une action en cours pour la redémarrer plus tard et de moduler les paramètres d'une action en train d'être exécutée par le Réaliseur. L'architecture ajoute de plus un ensemble de modules d'interprétation du comportement de l'utilisateur, la couche de Captation où sont implémentés les capteurs de l'agent, le module d'Interprétation du Comportement interprétant les actions multimodales de l'utilisateur et le module d'Interprétation de la Fonction Communicative interprétant les intentions communicatives de l'utilisateur. De plus, dans l'architecture ASAP, le contrôle du comportement de l'agent ne passe plus seulement par le Planificateur d'Intention, mais les résultats provenant des modules d'Interprétation du Comportement et de Captation peuvent directement être interprétés par le Planificateur de Comportement et le Réaliseur, permettant une coordination plus fine et plus rapide des actions de l'agent à son environnement, propriété essentielle pour l'implémentation de notre modèle.

L'architecture ASAP définit différents modules généraux permettant la perception et le contrôle du comportement de l'agent mais ne spécifie pas comment ces modules doivent être implémentés. Pour l'implémentation des modules, nous nous sommes donc inspirés de l'architecture informatique Ymir et son principe de perception et de décision distribuées entre plusieurs modules spécialisés dans la perception (percepteurs) et dans la décision (décideurs) (Thórisson, 2002). Le contrôle du comportement est ainsi distribué en couches de perception-action résultant d'une concurrence entre les décisions d'actions prises par plusieurs modules. Néanmoins, cette architecture est de nature événementielle, les décisions sont déclenchées à partir de règles. Dans notre implémentation, le contrôle des actions de l'agent est continu, nous avons donc dû adapter le mécanisme de percepteurs et de décideurs au contrôle continu du comportement.

4.2. Implémentation du modèle

Dans cette section nous décrivons en détail les différents modules permettant l'implémentation de notre modèle. La Figure 3 montre l'organisation générale de notre architecture.

4.2.1. Perception du comportement de l'utilisateur

En ce qui concerne la détection d'énergie acoustique et de hauteur de voix, nous nous sommes appuyés sur l'outil OpenSmile (Eyben *et al.*, 2013). À partir des valeurs d'énergie acoustique et de hauteur de voix fournies par OpenSmile, ces valeurs sont ensuite normalisées par les modules de captation d'énergie acoustique et de hauteur de voix afin d'avoir une valeur de signal comprise entre 0 et 1, manipulable par les équations de notre modèle. La formule permettant de normaliser les signaux est présentée équation (4).

$$s_n = \frac{s - s_{min}}{s_{max} - s_{min}} \quad (4)$$

s_n étant la valeur normalisée du signal de l'utilisateur (soit énergie acoustique soit hauteur de voix), s étant la valeur du signal provenant d'OpenSmile, s_{min} étant la valeur minimale du signal de l'utilisateur (récupéré grâce à une étape de calibration préliminaire) et s_{max} étant la valeur maximale du signal pour l'utilisateur. Une fois les valeurs normalisées calculées celles-ci sont envoyées aux modules de perception du comportement de l'agent.

Dans notre modèle la perception du comportement de l'utilisateur peut être décomposée en plusieurs processus distincts : le calcul du taux d'accumulation α (équation 2), le calcul de la valeur d'accumulation γ et la décision de changer ou non de rôle lorsque la valeur d'accumulation dépasse le seuil positif. Le calcul du taux d'accumulation et de la valeur d'accumulation est implémenté dans un ensemble de modules regroupé sur la figure 3 dans « Perception Prosodie ». Plus précisément, deux modules sont en charge de calculer, pour chaque signal reçu de l'utilisateur, les taux d'accumulation partiels α_j. Ces taux d'accumulation partiels

sont ensuite transmis au module se chargeant du calcul du taux d'accumulation α, selon l'équation (2) et enfin à un module se chargeant du calcul de la valeur d'accumulation γ. Le calcul du taux d'accumulation α est réalisé selon l'équation (2), celui de la valeur d'accumulation γ selon l'équation (1).

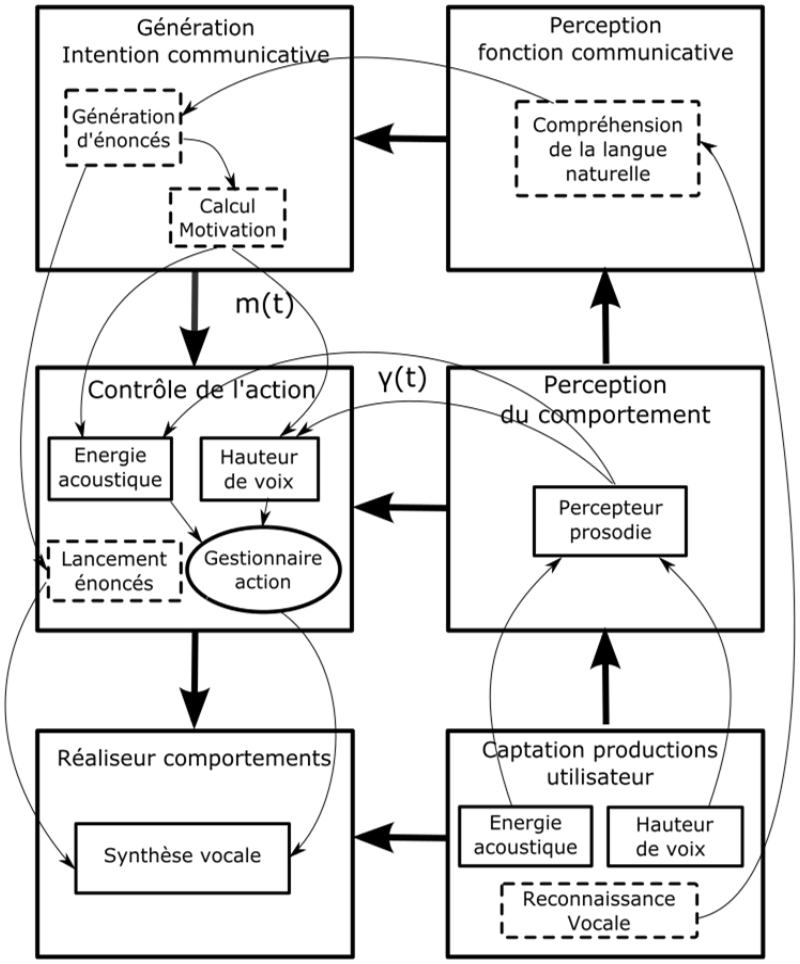

Figure 3. Schéma décrivant l'implémentation de notre modèle

4.2.2. Contrôle des actions

Les valeurs des signaux de l'agent sont contrôlées par les modules de contrôle de l'énergie acoustique et de hauteur de voix. Pour chaque module, l'équation spécifique du signal reprend l'équation générale définie par l'équation (3). Leur

implémentation spécifique correspond aux équations présentées dans (Jégou *et al.*, 2015).

Les variables d'actions modulées en sortie par les décideurs sont transmises au Gestionnaire d'Action sous forme de commandes d'action. C'est au Gestionnaire d'Action qu'appartient le rôle d'arbitrer d'éventuels conflits entre les commandes d'actions potentiellement contradictoires envoyées par différents modules de l'architecture. Le Gestionnaire d'Action se charge aussi de gérer l'exécution multimodale d'une commande d'action reçue des modules de contrôle des signaux. La conception du Gestionnaire d'Action est inspiré de l'Action Scheduler défini dans Ymir (Thórisson, 2002), en l'adaptant au contrôle d'actions continues.

4.2.3. Lien entre gestion du dialogue et gestion du tour de parole

Tel que présenté dans la section 4, l'implémentation de notre modèle nécessite une reconnaissance vocale et une synthèse vocale incrémentale. Avec une reconnaissance vocale incrémentale et l'utilisation d'un ensemble fini d'énoncés que l'agent est capable de reconnaître, l'agent peut formuler très tôt une hypothèse sur l'énoncé que l'utilisateur est en train de prononcer. 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.

La reconnaissance vocale incrémentale a été implémentée à l'aide de l'API Microsoft Speech. Le module de reconnaissance vocale fonctionne à l'aide d'une grammaire spécifiant l'ensemble des phrases de l'utilisateur que l'agent peut reconnaître. Lorsque l'utilisateur est en train de parler, le module de reconnaissance vocale fourni par l'API Microsoft Speech interprète et génère une hypothèse h sur la représentation sémantique de ce que l'utilisateur est en train de dire. Il y associe ensuite un degré de confiance c, renseignant le degré de probabilité que l'hypothèse représente bien l'énoncé prononcé par l'utilisateur. Cette hypothèse ainsi que le degré de confiance associé sont ensuite transmis aux percepteurs de l'architecture. Cette hypothèse est ensuite récupérée par le module de Compréhension de la langue naturelle qui évalue le degré de confiance c associée à l'hypothèse. Lorsque ce degré de confiance est supérieur à une valeur seuil, $c_{thre} = 0.8$, ce module considère que c est suffisamment grand pour que l'énoncé puisse être traité par le module de Génération d'Énoncés.

Une fois que le module de Génération d'Énoncés a déterminé l'énoncé que l'agent doit produire, cet énoncé est transmis au module de Lancement d'Énoncés, qui, en fonction, de contraintes associées à l'énoncé décide de transmettre ou non l'énoncé au Gestionnaire d'Action. Le gestionnaire d'action transforme la commande d'énoncé en commandes envoyées au réaliseur servant d'interface à la synthèse vocale.

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 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 d'énergie acoustique et de hauteur de voix provenant des requêtes de l'agent. Si la valeur d'énergie acoustique 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 d'énergie acoustique et de hauteur de voix. Il établit alors la correspondance entre la hauteur de voix et l'énergie acoustique, 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. Nous utilisons actuellement la synthèse vocale inpro_iSS (Baumann et Schlangen, 2012) pour moduler la prosodie de l'énoncé de l'agent.

Les modules de gestion du tour de parole et de génération de contenu verbal sont indépendants. La motivation à changer de rôle *m* est contrôlée par un module spécifique du module de Génération d'Intention. Dans ce module la motivation à changer de rôle dépend d'une part du fait que l'agent ait un énoncé ou non à prononcer, ainsi la motivation à changer de rôle est négative si l'agent est le locuteur courant, indiquant qu'il souhaite garder son rôle, et positive si l'agent est l'auditeur courant, indiquant qu'il souhaite prendre le tour. La valeur exacte de la motivation à changer de rôle est, elle, pour l'instant une constante que nous fixons manuellement. Dans des développements futurs de l'architecture, nous prévoyons d'inclure notamment des facteurs émotionnels dans le calcul de la motivation à changer de rôle.

Différents scénarios montrant la variabilité des stratégies de prise de parole de notre agent ont été reproduits par l'interaction entre notre agent et un utilisateur. Nous avons choisi la reconnaissance vocale SAPI de Microsoft pour l'interprétation incrémentale des énoncés de l'utilisateur. Chaque scénario correspond à une motivation m particulière de l'agent. Le module de génération de phrase implémenté utilise des paires adjacentes pour savoir quel énoncé produire en fonction de l'énoncé produit par l'utilisateur.

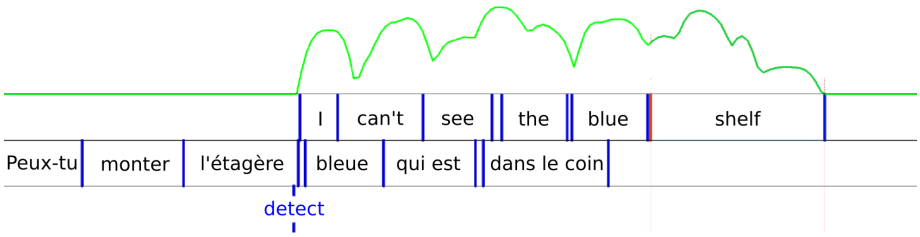

Figure 4 Scénario où l'agent a une motivation faible à parler. Dans ce cas de figure, bien que l'agent ait compris ce que l'utilisateur souhaitait dire, il attend que l'utilisateur ait fini de prononcer sa phrase pour commencer à parler

Nous illustrons les différents scénarios sur les figures 4 et 5. Sur chaque figure, les tours échangés entre l'agent (haut) et l'utilisateur (bas) sont retranscrits. Le moment où l'agent détecte la sémantique de l'énoncé de l'utilisateur et envoie sa réponse au Réaliseur est montré par l'annotation *detect.* Pour chaque exemple, nous

indiquons le profil d'énergie acoustique de l'agent, représenté par la courbe verte sur les figures, et nous montrons sous la figure la transcription du dialogue entre les deux participants.

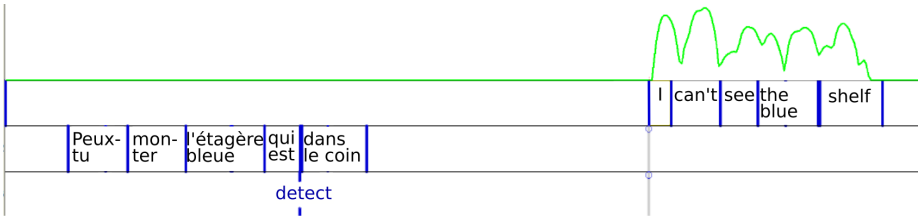

Figure 5 Scénario où l'agent a une motivation forte de parler. Dans ce cas de figure, il prend la parole en augmentant son énergie acoustique dès qu'il a compris ce que l'utilisateur voulait dire

La figure 4 montre un exemple d'interaction entre un utilisateur et un agent où l'agent reconnait l'énoncé de l'utilisateur avant que ce dernier ait fini de le prononcer. Une fois que l'agent a reconnu ce que l'utilisateur est en train de dire, la motivation forte de l'agent fait que ce dernier prend directement la parole sans attendre que l'utilisateur ait fini de parler. On observe un recouvrement long des deux tours de parole.

La figure 5 montre un cas de figure similaire : l'utilisateur prononce un énoncé détecté par l'agent avant que l'utilisateur ait fini son tour. Néanmoins, cette fois-ci l'agent a une motivation faible à parler, et, bien qu'il ait interprété et planifié ce qu'il comptait dire ensuite, il attend que l'utilisateur ait fini de parler avant de commencer à prendre le tour, ce qui résulte en une prise de tour après la fin du tour de l'utilisateur. On observe alors un moment de silence entre les deux tours.

5. Validation du modèle : interactions utilisateur-agent

5.1. Motivations

Nous présentons dans cette section une évaluation de l'interaction dialogique entre un agent piloté par notre modèle de gestion du tour de parole et un utilisateur. Le premier aspect de l'évaluation porte sur le ressenti de l'utilisateur en interaction avec notre agent, à savoir son sentiment d'aise dans l'interaction, ou encore la crédibilité de l'agent. Le second point évalué est la capacité de notre modèle à se coordonner avec l'utilisateur. Pour cela nous avons évalué le jugement de l'utilisateur sur la capacité de l'agent à prendre en compte les prises de parole de son interlocuteur. Lors des expérimentations, nous avons confronté les participants à des agents ayant des manières différentes de coordonner les tours de parole.

En général, les modèles de tour de parole existants sont évalués uniquement sur la base d'indicateurs de performance : les auteurs mesurent la capacité d'un agent conversationnel à prendre le tour en minimisant la durée de silence tout en évitant les chevauchements accidentels (Jonsdottir et Thórisson, 2013 ; Raux et Eskenazi, 2012). Ces approches supposent implicitement qu'une meilleure performance dans l'interaction avec le système conduit à un agent plus crédible et un meilleur engagement de l'utilisateur. Néanmoins, les conversations humaines sont par définition non optimales, la présence de coupures de parole et de pauses longues le montrant. Dans ce contexte d'interaction d'autres métriques de succès ou d'échec d'un modèle de tour de parole doivent être utilisées.

Certaines études ont ainsi directement évalué l'effet de différentes stratégies de tour de parole sur le ressenti de l'utilisateur. Ter Maat *et al.*, (2010) ont réalisé une expérimentation en magicien d'Oz, où l'utilisateur est chargé de répondre aux questions d'un agent. L'agent passe à la question suivante indépendamment de ce que dit l'utilisateur en variant ses stratégies de prise de parole : commencer à parler avant la fin de tour de l'utilisateur, commencer à parler juste après la fin de tour de l'utilisateur, et commencer à parler après avoir laissé une pause de quelques secondes. La perception que l'utilisateur a de l'agent est évaluée par l'intermédiaire d'un questionnaire comportant des questions liées à sa personnalité et ses compétences sociales. De Vault *et al.* (2015) et Skantze et Hjalmarsson (2010) proposent deux protocoles visant à évaluer des interactions dialogiques dans le cadre de scénarios d'initiative mixte, également en magicien d'Oz. L'interaction est évaluée à la fois par des questionnaires concernant la satisfaction de l'utilisateur, sa facilité à tenir la conversation, et la présence de pauses gênantes ou d'interruptions. Des mesures objectives liées aux durées de transition et aux coupures de parole sont de même réalisées et comparées entre les différentes conditions et avec des interactions humaines pour Skantze et Hjalmarsson (2010). Enfin, Cafaro *et al.* (2016) évaluent l'impression donnée par différents types d'interruptions sur le jugement de l'utilisateur concernant le caractère dominant ou soumis de l'agent.

Parmi les différents types d'interruptions testés, les auteurs montrent qu'une différence de perception peut être observée selon que l'agent effectue une interruption collaborative, complétant de manière collaborative l'énoncé du locuteur courant ou une interruption compétitive. Nous avons choisi de nous inspirer de ces études pour la conception de notre protocole expérimental. Ainsi, pour l'évaluation du ressenti de l'utilisateur nous avons varié la manière dont l'agent coordonnait ses tours de parole avec l'utilisateur. Dans un cas, l'agent laisse systématiquement la parole à l'utilisateur : s'il est locuteur, l'agent arrête l'énoncé en cours lorsque l'utilisateur cherche à l'interrompre, s'il est auditeur, il attend la fin de tour de l'utilisateur avant de prendre la parole. Dans l'autre cas, l'agent ne laisse jamais la parole à l'utilisateur s'il a quelque chose à dire : s'il est locuteur, il garde la parole si l'utilisateur cherche à lui couper la parole, s'il est auditeur, il saisit le tour son attendre la fin de tour de l'utilisateur. D'autres cas de figures auraient pu être testés, notamment le choix du type d'interruption réalisé de manière similaire à ce que Cafaro *et al.* (2016) a pu réaliser, néanmoins, la distinction entre interruption

collaborative ou compétitive étant en dehors du périmètre de notre modèle, nous avons choisi de nous focaliser sur ces deux conditions.

5.2. Protocole expérimental

Nous avons comparé 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 contrôlée par des règles simples où l'agent ne prend la décision de parler qu'après un intervalle de temps fixe après la fin de tour de l'utilisateur, et s'interrompt systématiquement lorsque l'utilisateur parle en même temps que lui, approche très souvent employée dans les architectures d'agent (Raux et Eskenazi, 2012) et considérée comme non-optimale. Nous reprenons les valeurs seuil utilisées dans la littérature (Ferrer *et al.*, 2002), 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 de coordination des tours de parole où l'agent varie sa motivation à changer de rôle 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 alors 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 où les prises de parole sont pilotées par notre modèle "condition 1", comprenant elle-même deux parties, la « condition 1 Weak », condition où l'agent a une motivation faible de parler et la « condition 1 Strong », condition où l'agent a une motivation forte. La condition correspondant au contrôle du tour de parole à base de temporisations est notée « condition 2 ».

Pour l'interaction dialogique utilisateur-agent, nous avons choisi un scénario de négociation que nous avions déjà utilisé pour l'analyse du tour de parole dans des conversations humaines. Dans ce scénario, on demande aux deux participants d'imaginer qu'ils sont sur un bateau en train de couler, et qu'ils se préparent à embarquer sur un bateau de sauvetage. Ils ont chacun des croyances différentes sur leur situation : l'un pense qu'il est proche de la côte, il souhaite donc favoriser des objets lui permettant de rejoindre rapidement la côte, l'autre participant pense qu'il est loin de la côte, il souhaite donc favoriser des objets lui permettant de survivre plusieurs semaines. Dans le scénario, il est suggéré au participant qu'il est proche de la côte, il lui est ainsi proposé trois objets à emporter sur le bateau de sauvetage. L'agent a, lui, la posture inverse. Il pense que le bateau est loin de la côte et souhaite donc favoriser trois autres objets. Afin de ne pas limiter les phrases que peut dire l'utilisateur pour être compris par l'agent, nous avons remplacé la composante 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 entre les deux modèles, il est nécessaire pour le magicien d'Oz de choisir la phrase à produire 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. Une partie de l'interface de choix des phrases est montrée sur la figure 6. Un clic sur le bouton au sommet génère une affirmation concernant la croyance de l'agent sur sa situation. Les deux boutons du dessous « Pouvoir se nourrir » et « Pouvoir se réchauffer » sont des affirmations concernant les stratégies que l'agent souhaite employer et les feuilles de l'arbre sont les objets résultants de ces stratégies. À chaque bouton est associé un ensemble de phrases de sorte que deux clics consécutifs sur le bouton ne génèrent pas la même phrase. Un arbre similaire est utilisé pour émettre des contre-arguments à l'utilisateur. L'ensemble des phrases générées par l'interface provient d'un corpus d'interactions humaines que nous avions recueilli au préalable selon le même scénario.

Figure 6. Partie de l'interface de choix des phrases prononcées par l'agent

Chaque participant interagit deux fois avec l'agent, chaque fois avec une condition différente et selon le même scénario. Chaque interaction dure deux minutes trente secondes. Afin d'éviter des effets d'ordre entre les passations, les conditions 1 et 2 ont été contrebalancées entre les participants. Nous n'avons néanmoins pas contrebalancé les conditions 1 Weak et les conditions 1 Strong dans la condition 1, le participant commençant toujours, dans la condition 1, par la condition 1 Weak, lui permettant de s'engager dans le dialogue puis finit par la condition 1 Strong où l'agent interrompt dès que possible l'utilisateur. Le choix de ne pas contrebalancer les conditions expérimentales provient de la volonté de ne pas introduire l'interaction entre l'utilisateur et l'agent par un scénario d'interaction où l'agent coupe systématiquement la parole à l'utilisateur pouvant mener à de fausses croyances concernant l'incapacité d'interagir avec l'agent et menant à un désengagement de l'utilisateur vis-à-vis de l'interaction. 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 de l'intentionnalité 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) et De Vault *et al.* (2015).

5.3. Résultats

31 étudiants, ingénieurs et chercheurs (30 hommes, une femme) ont participé à l'expérimentation. Tous avaient pour langue maternelle le français. Les résultats au questionnaire sont montrés sur le tableau 1. Peu de différences significatives ont été observées entre les conditions. Les participants ont globalement aimé parler avec l'agent (médiane de 7). Les résultats en matière 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 ont bien perçu que l'agent avait fait attention à ne pas leur couper la parole dans la condition 2 (médiane de 7.5) ce qui diffère significativement des réponses des participants dans la première condition (médiane de 4.5, degré de significativité $p = 0.006$). Néanmoins, les participants ont perçu contre intuitivement que l'agent leur a plus coupé involontairement la parole 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. Cela peut impliquer deux choses, soit l'agent a effectivement plus souvent détecté de manière erronée la fin de tour de l'utilisateur et a pris la parole en même temps que l'utilisateur était en train de parler, résultant en une « coupure de parole » involontaire dans la première condition, particulièrement la « condition 1 Weak » soit les participants n'ont pas perçu le caractère délibéré des coupures de parole de l'agent dans la « condition 1 Strong ». Nous avons analysé les occurrences de coupures de parole involontaires, c'est-à-dire les moments où l'agent a pris la parole en détectant de manière erronée la fin de tour de l'utilisateur. Le nombre de ces occurrences n'était pas significativement différents entre la condition 1 et la condition 2. Aussi, si l'utilisateur a perçu plus de coupures de parole involontaires, cela ne peut être dû qu'à la présence des interruptions délibérées de la « condition 1 Strong », qui ont été perçus comme involontaires par les participants.

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 transition utilisateur-agent est montrée sur la figure 7 et la répartition des durées de transition agent-utilisateur sur la figure 8. 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 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.

Tableau 1. Questions et réponses des participants pour les conditions 1 et 2

Questions	Médiane condition 1	Médiane condition 2	p-value
Q1 : « Ne percevait pas les moments où je parlais »	2.25	1.75	0.95
Q2 : « Prenait la parole aléatoirement »	2.5	2.625	0.6
Q3 : « M'a coupé la parole involontairement »	6	4	0.019*
Q4 : « M'a parfois délibérément coupé la parole »	6	6.5	0.91
Q5 : « A fait attention à ne pas me couper la parole »	4.5	7.5	0.006**
Q6 : « A mis du temps à me répondre »	3	2	0.77
Q7 : « A parfois refusé de me laisser parler »	6.125	5.75	0.16
Q8 : « J'ai été gêné par les prises de paroles de mon interlocuteur »	4.5	3.25	0.54
Q9 : « Je me suis senti à l'aise dans le dialogue »	5.25	6.25	0.55
Q10 : « J'ai aimé parler avec mon interlocuteur »	6.625	7	0.52
Q11 : « Le comportement de mon interlocuteur était proche d'un comportement humain »	5.625	6.5	0.97

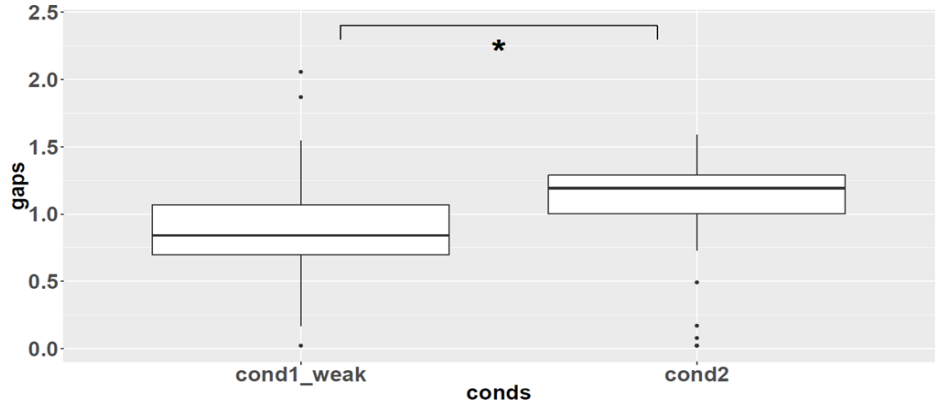

Figure 7. Répartition des durées de transitions utilisateur-agent

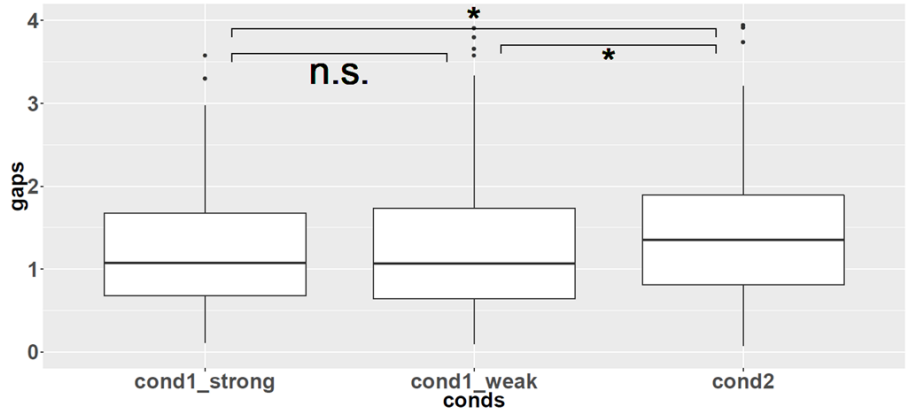

Figure 8. Répartition des durées de transition agent-utilisateur

5.4. 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 comportementales montrant une réaction de l'utilisateur : une hausse de hauteur de voix lors de la coupure de parole de l'agent, et une variation des temps de prise de parole de l'utilisateur. Nous avons, de plus, 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 « gênants » dans la conversation. Lorsque ces moments de silence étaient perçus, ils étaient considérés comme peu naturels, bien que deux participants aient 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 de nature audio, ne permettant pas à l'agent de fournir une rétroaction visuelle à l'utilisateur. Notons que la détection de la voix de l'utilisateur n'était pas parfaite : un nombre important de moments où aucune voix n'est détectée alors que l'utilisateur parle ont en effet été observés, provoquant un certain nombre de chevauchements et d'interruptions accidentels à la fin de tour de l'utilisateur. 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 de l'énergie acoustique de l'utilisateur.

6. Conclusion

Cet article apporte deux contributions. Premièrement, sur la base d'un modèle élaboré précédemment (Jégou *et al.*, 2015), nous proposons une architecture comportementale d'agent qui rend notre modèle opérationnel dans des scenarios de dialogue temps-réel. Notre modèle se distingue des modèles de contrôle du comportement d'un agent conversationnel traditionnels, événementiels. Notre modèle repose sur une perception continue des signaux de l'utilisateur et sur la modulation continue des signaux produits par l'agent. Peu d'architectures actuelles d'agents conversationnels permettent d'implémenter ce type de modèles. Pour intégrer notre modèle dans une architecture d'agent conversationnels, nous avons ainsi complété l'architecture ASAP (Kopp *et al.*, 2014) avec différents principes de

contrôle du comportement de l'agent inspiré de l'architecture Ymir (Thórisson, 1999). En utilisant l'architecture ASAP, respectant le standard SAIBA, nous permettons à notre modèle d'être intégrable avec un grand nombre de modules de contrôle du comportement d'un agent, gérant d'autres aspects des interactions utilisateur-agent, tels que l'interaction verbale, les émotions ou les attitudes interpersonnelles ou gérant d'autres signaux non-verbaux impliqués dans la coordination des tours de parole, tels que des réaliseurs gérant l'orientation du regard de l'agent.

Notre deuxième contribution est l'analyse de l'impact du mode de coordination de l'agent sur le ressenti de l'utilisateur. À titre de démonstration, nous avons montré la capacité d'un agent contrôlé par notre modèle à gérer la coordination des tours de parole dans des interactions dialogiques temps-réel avec des utilisateurs. Enfin, nous avons analysé des interactions temps-réel entre des utilisateurs et notre agent. Notre objectif était :

– de complémenter d'autres études perceptives évaluant l'expérience de l'utilisateur en interaction avec un agent modulant la manière dont il coordonne sa parole (De Vault *et al.*, 2015 ; Skantze & Hjalmarsson, 2010 ; Ter Maat *et al.*, 2010),

– de vérifier la capacité de notre agent à coordonner sa parole avec différents utilisateurs.

Nous avons ainsi mesuré différentes variables subjectives liées à la crédibilité de l'agent, sa présence sociale, la satisfaction et la facilité que l'utilisateur avait à interagir avec l'agent. Nous n'avons pas observé de différences dans les réponses des participants entre une condition où l'agent modulait la manière dont il coordonnait sa parole et une condition où l'agent suivait un ensemble de règles simples pour coordonner sa parole avec l'utilisateur. Nous avons aussi trouvé que les participants n'ont pas toujours perçu les interruptions de parole de l'agent comme volontaires.

Différents éléments pourraient être pris en compte pour améliorer l'interaction entre l'utilisateur et l'agent. Premièrement, le caractère volontaire ou involontaire des interruptions doit être évalué en lien avec le contenu de l'énoncé interrompu par l'agent : si l'interruption est effectuée trop tôt ou trop tard, il est probable que l'utilisateur perçoive l'interruption de l'agent comme inappropriée et peu crédible par rapport à une interruption compétitive réelle. Cela nécessiterait une extension de notre modèle pour modéliser la manière dont l'agent utilise le contenu verbal de l'énoncé de l'utilisateur pour coordonner sa parole. Deuxièmement, l'agent interagit actuellement avec l'utilisateur en modulant et interprétant uniquement l'énergie acoustique et la hauteur de voix. Ajouter des capacités d'interprétation multimodales améliorerait certainement la manière dont l'agent se coordonne avec l'utilisateur.

Bibliographie

Bailly G., & 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)*

Baumann T., & Schlangen D. (2012). INPRO_iSS: A Component for Just-in-time Incremental Speech Synthesis. In *Proceedings of the ACL 2012* System Demonstrations (p. 103–108).

Bevacqua E., Stanković I., Maatallaoui A., Nédélec A., & Loor P. D. (2014). Effects of Coupling in Human-Virtual Agent Body Interaction. In *Intelligent Virtual Agents 2014* (p. 54-63).

Cafaro A., Glas N., & Pelachaud C. (2016). The Effects of Interrupting Behavior on Interpersonal Attitude and Engagement in Dyadic Interactions. In *AAMAS 2016* (p. 911–920).

Cassell J., Bickmore T., Campbell L., & Vilhjálmsson H. (2000). Conversation as a System Framework: Designing Embodied Conversational Agents. *Embodied conversational agents*, 29–63.

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

De Vault D., Mell J., & Gratch J. (2015). Toward natural turn-taking in a virtual human negotiation agent. In *2015 AAAI Spring Symposium Series*.

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

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

Ferrer L., Shriberg E., & Stolcke A. (2002). Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody. In *ICSLP-2002* (p. 2061-2064).

Goldberg J. A. (1990). Interrupting the discourse on interruptions. *Journal of Pragmatics*, 14(6), 883-903.

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

Heldner M., & Edlund J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4), 555-568.

Huang L., Morency L.-P., & 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* – Vol. 3 (p. 1289–1290).

Jégou M., Lefebvre L., & Chevaillier P. (2015). A Continuous Model for the Management of Turn-Taking in User-Agent Spoken Interactions Based on the Variations of Prosodic Signals. In *Intelligent Virtual Agents 2015* (p. 389-398).

Jonsdottir G. R., & Thórisson K. R. (2013). A Distributed Architecture for Real-time Dialogue and On-task Learning of Efficient Co-operative Turn-taking. *Coverbal Synchrony in Human-Machine Interaction*, 293.

Kopp S., Krenn B., Marsella S., Marshall A. N., Pelachaud C., Pirker H., 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* (p. 205–217).

Kopp S., Welbergen H. van, Yaghoubzadeh R., & 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., & Wells B. (2013). Resources for turn competition in overlapping talk. *Speech Communication*, 55(5), 721-743.

Lessmann N., Kranstedt A., & 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* (p. 65).

Levitan R., Benus S., Gravano A., & Hirschberg J. (2015). Entrainment and Turn-Taking in Human-Human Dialogue. In *2015 AAAI Spring Symposium Series*.

McFarland D. H. (2001). Respiratory markers of conversational interaction. *Journal of Speech, Language, and Hearing Research: JSLHR*, 44(1), 128-143.

O'Connell D. C., Kowal S., & Kaltenbacher E. (1990). Turn-taking: A critical analysis of the research tradition. *Journal of Psycholinguistic Research*, 19(6), 345-373.

Oertel C., Wlodarczak M., Edlund J., Wagner P., & Gustafson J. (2013). Gaze patterns in turn-taking. In *13th Annual Conference of the International Speech Communication Association 2012 (INTERSPEECH 2012)*.

Ohshima N., Kimijima K., Yamato J., & Mukawa N. (2015). A conversational robot with vocal and bodily fillers for recovering from awkward silence at turn-takings (p. 325-330). *IEEE.*

Ratcliff R. (1978). A theory of memory retrieval. *Psychological review*, 85(2), 59-109.

Raux A., & Eskenazi M. (2012). Optimizing the Turn-taking Behavior of Task-oriented Spoken Dialog Systems. *ACM Trans. Speech Lang. Process.*, 9(1), 1:1–1, 23.

Ravenet B., Cafaro A., Biancardi B., Ochs M., & 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* (Vol. 9238, p. 375).

Roberts F., & Francis A. L. (2013). Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America*, 133(6).

Sacks H., Schegloff E. A., & Jefferson G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 696–735.

Schlangen D., Baumann T., Buschmeier H., Buss O., Kopp S., Skantze G., & 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* (p. 51–54).

Selfridge E., Arizmendi I., Heeman P., & Williams J. (2013). Continuously predicting and processing barge-in during a live spoken dialogue task. In *Proceedings of the SIGDIAL 2013 Conference* (p. 384–393).

Selfridge E. O., & Heeman P. A. (2009). A bidding approach to turn-taking. In *1st International Workshop on Spoken Dialogue Systems*.

Sellen A. J. (1995). Remote Conversations: The Effects of Mediating Talk with Technology. *Human-Computer Interaction*, *10*(4), 401-44.

Skantze G., & Hjalmarsson A. (2010). Towards incremental speech generation in dialogue systems. In *Proceedings of SIGDIAL 2010* (p. 1–8). Association for Computational Linguistics.

Ter Maat M., Truong K. P., & Heylen D. (2010). How turn-taking strategies influence users' impressions of an agent. In *Intelligent Virtual Agents* (p. 441–453).

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. & Thorisson H. T. (2010). A multiparty multimodal architecture for realtime turntaking. In *Intelligent Virtual Agents* (p. 350–356).

Warren W. H. (2006). The Dynamics of Perception and Action. *Psychological Review*, *113*(2), 358-389.

Wilson M. & Wilson T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review*, *12*(6), 957-968.