Boucles de perception-action et interaction face-à-face

Gérard Bailly, Frédéric Elisei & Stephan Raidt
Département Parole & Cognition, GIPSA-lab, UMR 5216
CNRS/Grenoble Universités

Résumé

Cet article explore un champ de recherches en plein essor : la communication face-à-face. Les performances et la robustesse des composants technologiques nécessaires à la mise en œuvre de systèmes d'interaction face-à-face entre l'homme et un agent conversationnel – technologies vocales, vision par ordinateur, synthèse d'images, compréhension et génération de dialogues, etc. – sont maintenant matures. Nous esquissons ici un programme de recherche centré sur la modélisation des diverses boucles de perception-action impliquées dans la gestion de l'interaction et sur le paramétrage dynamique de ces boucles par les divers niveaux de compréhension de la scène dans laquelle humains, robots et agents conversationnels animés seront inévitablement plongés.

1 Introduction

L'essentiel des technologies vocales et des connaissances phonétiques que nous avons sur les langues orales ont été acquises et développées grâce à l'enregistrement et l'étude des signaux audiovisuels de la communication langagière. Il est vrai que, sans en avoir véritablement conscience, nous parlons au quotidien avec des microphones et des caméras qui semblent devenus de manière naturelle des interlocuteurs aussi légitimes que nos alter-egos. Pourtant l'essentiel de nos productions langagières et de nos interactions sociales est composé de communications face-à-face, situées, en prise directe avec notre environnement physique immédiat. De même, l'émergence de nouvelles technologies opportunistes, ubiquitaires, centrées sur l'utilisateur, de services géolocalisés, nous pousse à réintégrer les dimensions interactionnelles et pragmatiques dans l'étude de la parole et de l'organisation du discours, conçu non plus comme un simple encodage d'informations linguistiques et paralinguistiques mais comme le moyen privilégié d'agir sur le monde et d'interagir avec les agents humains ou artificiels qui le peuplent.

Ces interactions sont multimodales : si la parole est un vecteur d'action privilégié, elle est accompagnée et complémentée par de multiples autres modalités gestuelles (expressions faciales, mouvements de tête, posture, gestes brachio-manuels, regard, etc.). La plupart de ces gestes s'inscrivent dans la même structuration linguistique que la parole que ce soit pour renforcer la mise en relief ou la valence d'un élément de discours (mouvements de sourcils dans Cavé, Guaïtella et al. 2002; et Flecha-García 2004), pour gérer les tours de parole (Edlund, Heldner et al. 2005), planifier le discours (Goldin-Meadow, Nusbaum et al. 2001) ou pour complémenter l'action (Louwerse and Bangerter 2005).

Cet article présente le cadre général de modélisation de l'interaction que nous sommes en train de mettre en œuvre afin de doter un agent conversationnel de capacités à établir une interaction face-à-face située avec un partenaire humain. Les compétences attendues des ingrédients technologiques nécessaires à la mise en œuvre d'un tel système et les modèles du comportement humain qu'ils exploitent sont illustrées par les recherches que nous menons depuis 5 ans sur la gestion du regard et sa relation à l'activité langagière.

2 Boucles de perception-action et systèmes d'interaction

Les boucles de perception-action qui conditionnent la planification de nos actions de communication prennent en compte de nombreux paramètres extraits de notre analyse de l'environnement à des échelles de temps – donc des temps de réaction - très différentes et des niveaux de traitement cognitif hétérogènes allant de simples boucles réflexes jusqu'à la gestion plus complexe de rapports sociaux et culturels.

Le modèle d'interaction que nous nous proposons de paramétrer avec des données comportementales et dont voulons doter un agent conversationnel est présenté Figure 1. Il est largement inspiré du modèle YMIR proposé par Thörisson (Thórisson 2002) et consiste essentiellement en un couplage de boucles de perception-action, conditionnant les actions de communication du locuteur par les traitements de divers niveaux d'analyse et de compréhension de la scène d'interaction dans laquelle ce dernier est plongé. Ce modèle distingue essentiellement trois boucles dont les temps de réaction sont de plus en plus lents car travaillant sur des unités de discours de plus en plus larges : (a) un niveau dit réactif ou réflexe qui réagit de manière instinctive aux sollicitations de l'environnement ; ce niveau intègre la plupart des mécanismes attentionnels (attention visuelle, ajustement du volume sonore, retours de canal, convergences multimodales, etc.) permettant à l'organisme de s'adapter de manière quasi-automatique aux signaux émis par l'environnement immédiat ; (b) un niveau dit cognitif, gérant la communication dans son ensemble, aussi bien le contenu linguistique et émotionnel des énoncés que la gestion des tours de parole et des signaux signalant la mise en œuvre effective des processus d'élaboration de cette communication et d'estimation de l'état cognitif de l'interlocuteur ; (c) un niveau dit de profilage, qui élabore un paramétrage des deux autres niveaux en fonction des intentions de communication, du profil psychologique, social et affectif du locuteur et de son estimation du profil et intentions de l'interlocuteur. Ces trois niveaux se conditionnent évidemment les uns les autres puisque l'analyse des actions et réactions de l'interlocuteur actualisent ces divers niveaux d'analyse de la situation de communication.

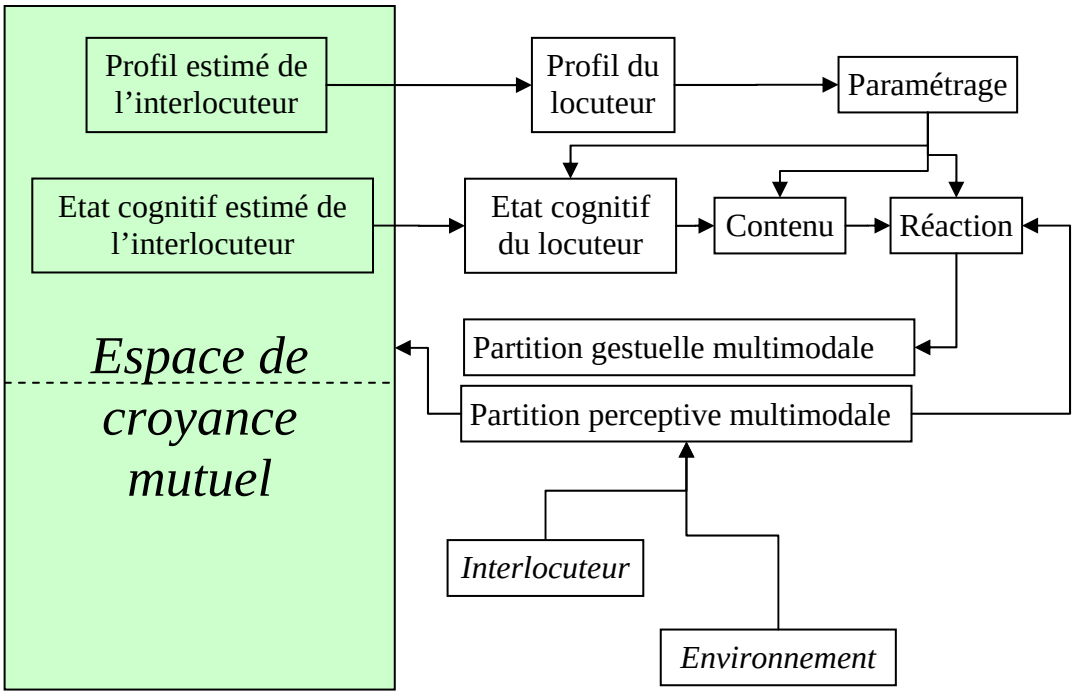

Figure 1 : Schéma général de contrôle des boucles d'interaction conditionnant la réponse du locuteur aux changements de son environnement survenant de sa propre volonté ou de l'interlocuteur.

La suite de cet exposé démontre l'intérêt de ces divers niveaux d'analyse et de compréhension de la scène en s'appuyant sur nos recherches sur la gestion du regard en situation de communication face-à-face suggérant des études similaires sur le signal de parole.

Figure 2 : agent conversationnel sensible à l'environnement. A gauche, à l'apparition d'un autre interlocuteur dans la scène ; à droite : à un objet agité devant son nez. (d'après Picot, Bailly et al. 2007)

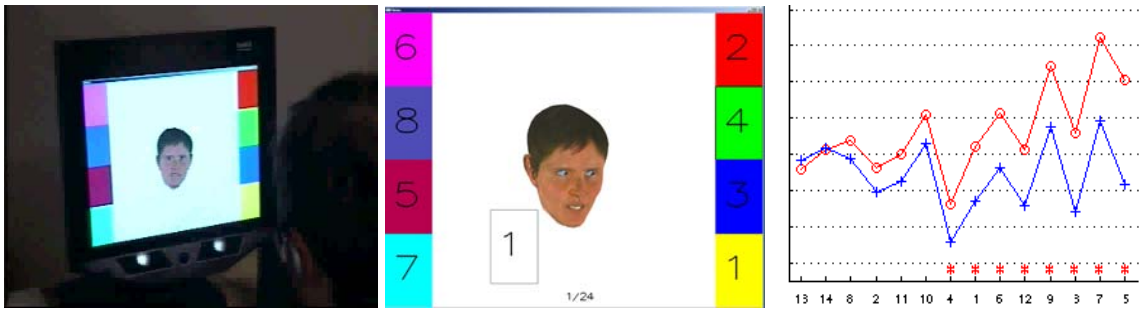

Figure 3 : agent conversationnel attirant l'attention d'un interlocuteur sur un objet de l'environnement – virtuel en l'occurrence. A gauche, le dispositif utilisant un oculomètre afin d'exploiter et enregistrer le regard de l'usager ; au centre, l'agent désigne le lieu où poser la carte tirée dans un jeu d'appariement ; à droite : les durées d'appariement (une ligne horizontale = 100 ms) sont significativement augmentées pour plus de la moitié des sujets lorsque l'agent désigne des lieux inadéquats et ceci bien que les sujets soient instruits de ne pas tenir compte des actions de l'agent. (d'après Raidt, Bailly et al. 2006)

3 Boucles réactives

3.1 Attention visuelle et auditive

De nombreux systèmes d'attention visuelle (voir notamment la description de l'un des plus cités ds Itti, Dhavale et al. 2003) ont été proposés afin de rendre compte de nos stratégies de scrutation d'une scène. Ces systèmes se composent essentiellement de trois modules : (a) un premier module simulant le traitement de l'image rétinienne par l'ensemble des cônes et bâtonnets. Ce module calcule ainsi généralement plusieurs cartes - dites de saillance - sensibles au mouvement, à la saturation des canaux de couleur ou à divers gradients de luminosité suivant des directions privilégiées – dont notamment verticales et horizontales peuplant notre environnement soumis à la gravité. Les points les plus saillants de ces cartes constituent autant de candidats à une fixation oculaire par notre œil directeur. (b) une carte dite de pertinence est calculée afin de moduler ces saillances par certaines caractéristiques de la tâche ; ce mécanisme d'attention visuelle descendante permet ainsi d'expliquer certaines cécités attentionnelles (Simons and Chabris 1999) où un objet intrinsèquement saillant par son mouvement ou sa taille mais n'ayant pas les caractéristiques d'apparence recherchées n'est pas détecté. (c) le module d'attention proprement dit gère le déclenchement des saccades oculaires, la durée des fixations – déclenchant notamment des processus de plus haut niveau tels que la détection, l'identification d'objets ou de visages et, le cas échéant, la gestion de l'interaction face-à-face et la gestion de la poursuite d'éléments de la scène en mouvement. Les modèles diffèrent dans les priorités données aux composantes de l'analyse de scène –

segmentation, principes de regroupement et de décomposition des sous-éléments constitutifs d'un objet ou corps rigide ou articulé (Sun 2003), etc. - et la gestion dynamique de la pertinence des points d'intérêt – pile d'attention (Picot, Bailly et al. 2007), système d'alerte, etc.

Ce système d'attention visuelle, nous rendant intrinsèquement attentifs et conscients du monde physique qui nous entoure (cf. Figure 2), permet aux niveaux supérieurs de collecter des informations essentielles sur les objets et agents de l'environnement : position, vitesse, nature, etc. Ces niveaux supérieurs vont pouvoir à leur tour utiliser ces informations pour ébaucher une compréhension de la scène qui s'enrichira par l'interaction : position et identification des objets et agents de l'environnement (devenant ainsi autant de référents potentiels du discours), caractéristiques individuelles (poids, sexe, âge, etc.), activité cognitive des interlocuteurs… notamment au travers de l'observation de leurs propres stratégies de scrutation de la scène.

Comme bien d'autres espèces animales dotées d'un système de vision, nous sommes ainsi très sensibles au regard de l'autre, posé sur nous ou sur un autre objet d'intérêt. De nombreuses études ont montré que le regard est une des composantes essentielles de l'attention sociale : nous ne pouvons échapper au biais d'attention imposé par le regard de l'autre (Langton and Bruce 1999; Langton, Watt et al. 2000) et nous avons montré que ce biais peut être reproduit par des agents conversationnels incarnés (cf. Figure 3), d'autant plus efficacement que cette monstration est multimodale et s'accompagne d'une désignation brachio-manuelle ou vocale (Raidt, Bailly et al. 2006).

Nous reviendrons plus loin sur les termes de ce couplage entre jeux de regard et sur l'influence que les niveaux cognitifs et sociaux exercent sur leur paramétrage.

3.2 Convergence… ou pas

Cette attraction mutuelle n'est pas l'apanage exclusif des jeux de regards. La voix est aussi un terme de l'attention sociale : des études récentes montrent que les boucles de perception-action peuvent tendre les interlocuteurs non seulement à faire converger leurs productions linguistiques vers un répertoire de formes communes – lexique, syntaxe, etc. - (Gutierrez-Clellen and Heinrichs-Ramos 1993) mais aussi à adapter mutuellement leurs productions vocales – formes phonologiques telles que le choix des variantes de prononciation, etc. mais aussi caractéristiques phonétiques telles que la vitesse d'articulation, les formants vocaliques, la distribution de la fréquence fondamentale, ou plus prosaïquement le niveau sonore (Ward and Nakagawa 2004; Komatsu and Morikawa 2005; Kousidis, Dorran et al. 2008). Ces convergences ont été observées aussi bien entre interlocuteurs humains que dans des dialogues homme-machine : les traces collectées lors de dialogues d'agents conversationnels avec des usagers montrent que ces derniers adaptent très rapidement leur lexique et la syntaxe de leurs interventions au niveau de langue utilisée par l'agent (Bell, Gustafson et al. 2003). En retour, l'adaptation des productions vocales des agents conversationnels à celles de l'usager est appréciée par les usagers (Zoltan-Ford 1991). Les systèmes d'interaction disposent donc ainsi d'un outil précieux de contrôle de la boucle d'interaction par l'action afin de faciliter le traitement de la scène multimodale par leur système de perception.

Ce couplage réactif est évidemment perméable aux impératifs cognitifs et sociaux : ainsi ces tendances générales à l'unisson sont largement modulées par les contraintes linguistiques de la langue utilisée et peuvent aussi être inhibées par le code social : ainsi dans son étude des productions vocales de japonais en milieu domestique, Campbell (Campbell 2006) montre notamment que la structure familiale se reflète dans l'usage de codes fréquentiels et de qualités de voix adaptés aux rapports sociaux établis au sein de la famille japonaise.

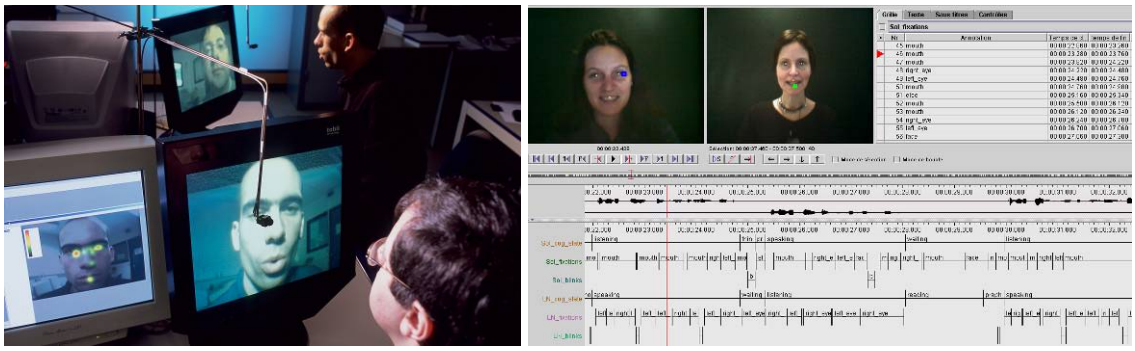

Figure 4 : à gauche, le dispositif de caméras, écrans et oculomètres croisés permettant de mesurer les regards des interlocuteurs en communication face-à-face ; à droite, la vérification de l'étiquetage automatique des durées de fixations et des états cognitifs.

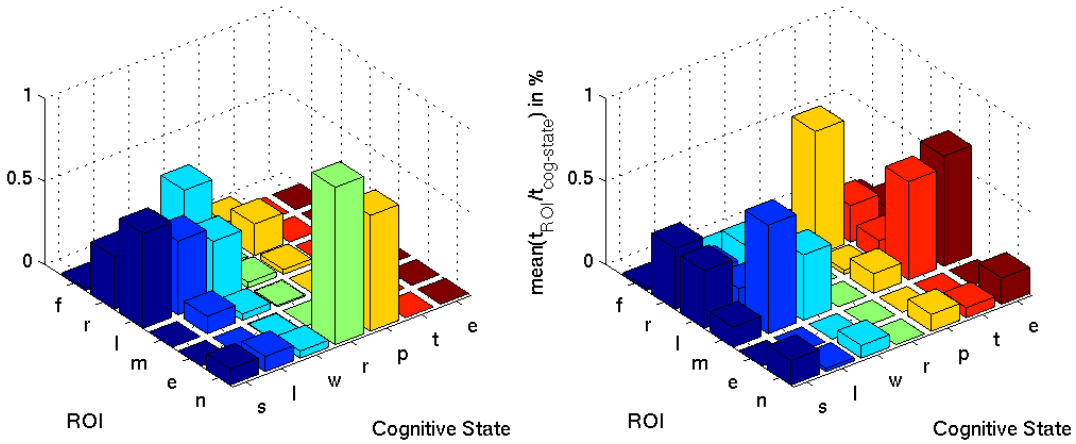

Figure 5 : probabilité d'observer une fixation sur une région du visage (f=face, r= œil droit, l=œil gauche, m=bouche, e=écran, n=ailleurs) étant donné un état cognitif impliqué dans l'interaction (s : parler, l=écouter, w=attendre, r=lire, p=préparation à parler, t=réfléchir, e=autre). A gauche, caractéristiques de notre locutrice cible lorsqu'elle initie l'échange, à droite : lorsqu'elle subit l'échange.

4 Boucles cognitives

Le face-à-face est donc mis en place, les partenaires sont attentifs l'un à l'autre et l'échange d'informations, la construction d'un espace d'expériences et de croyances mutuelles va pouvoir s'engager ou se poursuivre.

Au-delà de l'analyse et la construction de sens en ligne, engendrant la richesse et la complexité de la production de parole spontanée que ce soit au niveau lexical, syntaxique, sémantique, phonologique et phonétique, les indices perceptifs permettant aux interlocuteurs de réguler les tours de parole, d'estimer ce que l'autre ressent sur ce qu'il entend ou dit, constituent un puzzle scientifique qui mérite au moins autant d'attention que l'analyse et la planification du discours lui-même. Dans l'interaction, la compréhension des productions verbales de l'autre est aussi importante que la planification de nos productions verbales, productions qui, en retour, reflètent nos attentes et notre stratégie de découverte des motivations de l'autre. L'écoute est active et livre des indices sur les processus de compréhension mis en jeu… qui influencent en retour nos stratégies de production. L'écoute s'accompagne ainsi souvent de clavardage (stigmates visibles ou audibles de la gestion de l'interaction par l'interlocuteur ou back-channeling). Ces stigmates sont multimodaux : regard, expressions faciales, posture et parole participent à cette gestion de l'interaction. Ces murmures et onomatopées sont d'une richesse incomparable et constituent une part importante des productions vocales des interlocuteurs (Campbell 2004), modulées par un nombre tout aussi important de formes prosodiques que le discours qu'elles ponctuent.

Nous avons montré (Raidt, Bailly et al. 2007) que les états cognitifs avaient un impact important sur les jeux de regard observés en situation de communication face-à-face médiatisée (cf. Figure 5). L'attention visuelle reflète les attentes des interlocuteurs sur les zones du visage où l'information est susceptible d'être lue : la bouche est plus scrutée lorsqu'on attend ou écoute une information divulguée imprédictible, on attend un contact visuel pour engager la production de parole, etc. Le clignement des yeux, geste régulier (plus de 10000 clignements des yeux par jour) permettant d'humidifier la cornée, est bien sûr sensible à l'environnement (lumière, particules en suspension, etc.) mais aussi à l'état physiologique (fatigue, etc.) et cognitif du sujet : nous clignons plus lorsque nous parlons que lorsque nous écoutons (Bailly, Elisei et al. 2006). De même, certains gestes de mains (iconiques, déictiques), certaines expressions faciales (dégout, tristesse) sont plus susceptibles d'être déclenchés lors de la production de parole que lors de la phase de réflexion préalable ou lors des phases de perception de la parole de l'autre.

5 Paramétrage individuel et interaction sociale

Les comportements décrits plus haut sont des comportements moyens, signant de manière significative les divers évènements internes et externes impliqués dans l'interaction dans laquelle les interlocuteurs sont plongés. Ces comportements exhibent une large variabilité et, au-delà des codes et normes sociales qui régissent et paramètrent nos comportements, on constate dans toutes les études une forte variabilité inter- et intra-locuteurs. C'est pourquoi nous choisissons dans nos études de confronter un locuteur-cible à de multiples scénarios d'interaction, impliquant des interlocuteurs variés. Même dans ce cadre très contraint, il est important de contrôler ou de caractériser les partenaires de l'interaction observée : le sexe, l'âge relatif des interlocuteurs, leurs positions sociales, le degré d'intimité sont autant de paramètres influant sur les boucles de perception-action.

Dans notre étude sur les jeux de regard, nous avons choisi des interlocutrices familières, d'âge comparable de même niveau social. Si le rôle tenu par chaque interlocutrice dans la conversation a un impact significatif sur la distribution du regard sur les parties du visage, le facteur « interlocuteur » est significatif et la personnalité (intra/extravertie, etc.) de chaque interlocuteur a un fort impact sur les stratégies attentionnelles.

Le paramétrage social interagit fortement avec les composants de bas niveau de l'analyse de scène. Nous sommes notamment très sensibles à la distance de communication : Hall (Hall 1963) propose 4 cercles égocentrés délimitant des espaces dans lesquels des interactions peuvent ou ne peuvent pas se dérouler harmonieusement. Kismet, le robot affectif développé au MIT [], gère avec une boucle de perception-action très simple son engagement ou désengagement de l'interaction dans son cercle intime en fonction de son empathie et de l'état affectif estimé de son interlocuteur. La voix - son intensité mais aussi le régime de phonation, le timbre – est affectée par la gestion de cet espace interpersonnel et constitue un instrument irremplaçable de gestion de l'engagement du locuteur dans l'interaction. On voit donc que l'estimation de paramètres physiques telles que l'intensité du bruit ambiant ou la distance inter-personnelle a une incidence importante sur la gestion de l'interaction et la nature des informations confidentielles ou non que nous pouvons être amenés à divulguer.

6 Conclusions

Nous avons essayé de montrer que la parole fait partie d'un arsenal multimodal de gestion de l'interaction et qu'il est difficile de restreindre l'étude de la parole à son seul rôle d'encodage de l'information linguistique voire paralinguistique. Même en restreignant son rôle à ces fonctions, la situation d'interaction, l'importance du contenu linguistique, les relations sociales établies ou supposées entre interlocuteurs ont un impact déterminant sur cet

encodage. Lors d'une conversation face-à-face, la phrase n'est pas une bouteille jetée à la mer, une lettre de la dernière chance, c'est un moyen d'action sur le monde, une progression infime dans la compréhension de notre environnement et un de nos moyens d'influer sur le processus dynamique en cours.

L'étude des systèmes dynamiques humains est difficile, en grande partie parce qu'elle heurte la démarche scientifique qui s'appuie sur la répétabilité des expériences, mêmes causes – mêmes effets. Chaque interaction est unique. Pour satisfaire à la démarche scientifique, il faut que nous puissions progresser dans la compréhension des divers constituants et couplages entre boucles de perception-action grâce à des expériences contrôlées permettant de maintenir un niveau de complexité acceptable tout en permettant de recueillir des données pouvant être soumises à l'analyse et la modélisation statistique. Nous devons renouveler nos dispositifs de métrologie – l'observation biaisée par l'observateur se substituant trop souvent à la mesure – afin de caractériser les comportements humains multimodaux dans des situations d'interaction face-à-face réalistes et finalisées. Nous avons utilisé des paradigmes de jeux – jeux de langage, jeux de cartes, construction collaborative de pièces, etc. – afin de pouvoir également mesurer l'impact sur le déroulement de l'interaction des modèles de comportement dont nous dotons nos agents conversationnels animés. La phase d'évaluation subjective et objective de la qualité de l'interaction fait en soi partie de la démarche scientifique et permet de reboucler sur l'expérimentation et la modélisation. Cette phase d'évaluation doit s'appuyer sur la mise en œuvre de systèmes d'interaction simulés en temps-réel. Cette contrainte a un impact fort sur les développements technologiques : la course à la robustesse et au taux de reconnaissance devra céder la place à des systèmes multimodaux plus complexes s'appuyant sur des technologies élémentaires simples, sélectives, partiales et peu fiables… à l'image de notre système cognitif.

Remerciements

Cette réflexion s'appuie sur un travail d'équipe. Nous tenons donc à remercier les nombreux étudiants qui ont contribué à l'ébauche de ce projet scientifique : Alex Bartroli, Maxime Bérar, Alix Casari, Maxime Mantovani, Antoine Picot & Romain Rossi. Cette réflexion s'appuie aussi sur les échanges avec de nombreux chercheurs qui partagent cette vision, dont Rachid Alami, Nick Campbell, Peter F. Dominey et James L. Crowley. Il résulte des travaux conduits au sein de nombreux projets dont le projet Présence financé par le Cluster ISLE de la région Rhône-Alpes, le Plan Pluri-Formations « Interactions Multimodales » financé par les 4 universités de Grenoble et le projet AMORCES financé par l'ANR.

Références

Bailly, G., F. Elisei, et al. (2006). Embodied conversational agents : computing and rendering realistic gaze patterns. Pacific Rim Conference on Multimedia Processing, Hangzhou.

Bell, L., J. Gustafson, et al. (2003). Prosodic adaptation in human-computer interaction. International Congress of Phonetic Sciences, Barcelona.

Campbell, N. (2004). Listening between the lines; a study of paralinguistic information carried by tone-of-voice. International Symposium on Tonal Aspects of Languages: Emphasis on Tone Languages, Beijing.

Campbell, N. (2006). On the structure of spoken language. Speech Prosody, Dresden, Germany.

Cavé, C., I. Guaïtella, et al. (2002). Eyebrow movements and voice variations in dialogue situations: an experimental investigation. International Conference on Spoken Language Processing, Denver, CO.

Edlund, J., M. Heldner, et al. (2005). Utterance segmentation and turn-taking in spoken dialogue systems. Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen. B. Fisseni, H.-C. Schmitz, B. Schröder and P. Wagner. Frankfurt am Main, Germany, Peter Lang**:** 576-587.

Flecha-García, M. L. (2004). Eyebrow raising and communication in map task dialogues. Conference of the European Society for Cognitive Psychology, Grenada - Spain.

Goldin-Meadow, S., H. Nusbaum, et al. (2001). "Explaining math: gesturing lightens the load." Psychological Sciences **12**(6): 516-522.

Gutierrez-Clellen, V. F. and L. Heinrichs-Ramos (1993). "Referential cohesion in the narratives of Spanish-speaking children. A developmental study." Journal of Speech and Hearing Research **36**: 559-567.

Hall, E. T. (1963). "A system for the notation of proxemic behaviour." American Anthropologist **85**: 1003–1026.

Itti, L., N. Dhavale, et al. (2003). Realistic avatar eye and head animation using a neurobiological model of visual attention. SPIE 48th Annual International Symposium on Optical Science and Technology, San Diego, CA.

Komatsu, T. and K. Morikawa (2005). Entrainment of rate of utterances in speech dialogs between users and an auto response system. Knowledge-Based Intelligent Information and Engineering Systems (KES), Melbourne, Australia.

Kousidis, S., D. Dorran, et al. (2008). Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues. Interspeech, Brisbane.

Langton, S. and V. Bruce (1999). "Reflexive visual orienting in response to the social attention of others." Visual Cognition **6**(5): 541-567.

Langton, S., J. Watt, et al. (2000). "Do the eyes have it ? Cues to the direction of social attention." Trends in Cognitive Sciences **4**(2): 50-59.

Louwerse, M. M. and A. Bangerter (2005). Focusing attention with deictic gestures and linguistic expressions. Annual Conference of the Cognitive Science Society (CogSci), Stresa, Italy.

Picot, A., G. Bailly, et al. (2007). Scrutinizing natural scenes: controlling the gaze of an embodied conversational agent. International Conference on Intelligent Virtual Agents (IVA), Paris.

Raidt, S., G. Bailly, et al. (2006). Does a virtual talking face generate proper multimodal cues to draw user's attention towards interest points? Language Ressources and Evaluation Conference (LREC), Genova, Italy.

Raidt, S., G. Bailly, et al. (2007). Gaze patterns during face-to-face interaction. IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshop on Communication between Human and Artificial Agents (CHAA), Fremont, CA.

Simons, D. J. and C. F. Chabris (1999). "Gorillas in our midst: sustained inattentional blindness for dynamic events." Perception **28**(9): 1059-1074.

Sun, Y. (2003). Hierarchical object-based visual attention for machine vision. Institute of Perception, Action and Behaviour. School of Informatics. Edinburgh, University of Edinburgh**:** 169.

Thórisson, K. (2002). Natural turn-taking needs no manual: computational theory and model from perception to action. Multimodality in language and speech systems. B. Granström, D. House and I. Karlsson. Dordrecht, The Netherlands, Kluwer Academic**:** 173–207.

Ward, N. and S. Nakagawa (2004). "Automatic user-adaptive speaking rate selection." International Journal of Speech Technology: 259-268.

Zoltan-Ford, E. (1991). "How to get people to say and type what computers can understand." International Journal of Man-Machine Studies **34**: 527-547.