Des machines parlantes aux agents conversationnels incarnés

Gérard Bailly, Frédéric Elisei, Stephan Raidt

► To cite this version:

Gérard Bailly, Frédéric Elisei, Stephan Raidt. Des machines parlantes aux agents conversationnels incarnés. Catherine Garbay, Daniel Kayser (Eds.). Informatique et Sciences Cognitives : influences ou confluences?, Ophrys, pp.215-234, 2011. hal-00634982

HAL Id: hal-00634982

<https://hal.science/hal-00634982>

Submitted on 24 Oct 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Des machines parlantes aux agents conversationnels incarnés

G. Bailly, F. Elisei et S. Raidt

Introduction

Le super-robot Astroboy, imaginé par Osamu Tezuka en 1951, a l'apparence d'un enfant et vit en bonne société avec les hommes. Rejeté par son père créateur, il est recueilli par un savant qui en fait un super-héros combattant pour la paix, la justice et la tolérance. Ce manga est symptomatique de la conception japonaise d'une robotique apprivoisée, socialisée et pacifique, vivant en harmonie avec l'humanité (voir Kaplan 2006 pour une analyse plus complète). Deux mois après la naissance supposée d'Astroboy (le 7 avril 2003 !), ATR a créé un laboratoire de neuroscience computationnelle et confié ses recherches en robotique anthropoïde à des spécialistes en **neurosciences**, neurobiologie et imagerie cérébrale.

Cette confluence de l'informatique avec la biologie et les sciences humaines et sociales est effectivement révélatrice des besoins de la robotique anthropoïde sociale, challenge ultime de la biocybernétique. Le robot en interaction avec un environnement réel doit en effet piocher dans cet environnement une part non négligeable des paramètres des multiples tâches qu'il est en mesure d'exécuter. Lorsque cet environnement comprend des partenaires humains, le robot ne possède alors qu'une faible partie de la spécification de la tâche, l'autre étant entre les mains de ses partenaires de communication et échappant de fait pour partie à ses concepteurs. Une solution adéquate doit donc être négociée et l'efficacité de la robotique passe par

la socialisation, la communication et l'attention mutuelle entre agents humains et artificiels, engageant ainsi une reconstruction artificielle de nos échanges humains. L'activité des partenaires humains offre effectivement à qui sait les observer les symptômes de la solution calculée par l'intelligence humaine **qui lui fait face**. Intégrer l'humain dans la boucle de contrôle présente donc l'avantage de pouvoir piocher et exploiter ces symptômes.

Les règles de communication que ces robots doivent adopter sont donc les nôtres. De la compréhension du langage articulé aux gestes iconiques ou déictiques, ces règles impliquent toutes les dimensions sociales, économiques, techniques, locales, spirituelles, morales qui fondent notre structure sociale et façonnent notre état mental, psychologique et affectif.

Il faut non seulement doter ces partenaires artificiels de capacités à percevoir et à interroger le monde qui les entoure mais aussi de capacités à signaler aux partenaires que ce processus a effectivement lieu, que ces robots analysent, cherchent à comprendre nos intentions et développent *in fine* une véritable théorie de l'esprit (Povinelli and Preuss 1995) leur permettant de s'insérer dans la collectivité humaine. Si ces représentations mentales doivent s'appuyer sur une analyse de scène multimodale¹ très complète, les gages effectifs de la réalité de ces représentations, dont ces créatures artificielles doivent nous convaincre, sont de leur côté tout aussi indispensables : par exemple, indiquer à l'interlocuteur qu'on a détecté sa présence en se plaçant face à lui et en dirigeant son regard vers son visage. En adoptant nos comportements, nos attitudes, nos règles de gestion de l'interaction et de la conversation humaine, la génération de formes d'action adéquates contribuent non seulement à la mise en présence du traitement de l'information mais aussi à sa crédibilité et sa robustesse.

Si les robots humanoïdes exacerbent nos fantasmes de créateurs depuis l'aube de l'humanité, d'autres créatures virtuelles engendrées par l'informatique moderne sont venues grossir les rangs des agents

¹ C'est-à-dire combinant l'ensemble des modalités motrices et sensorielles disponibles.

sociaux. Introduits par Cassell et al. (2000), les agents conversationnels incarnés (« embodied conversational agents ») ou plus simplement animés (« agents conversationnels animés » ou ACA) sont des créatures virtuelles (affichées sur un écran) capables d'engager la conversation avec des partenaires humains en comprenant et produisant la parole ainsi que, de manière plus large, les gestes corporels (gestes de main, expressions faciales, etc.) caractéristiques du comportement humain. Comparée aux robots anthropoïdes, la mise en présence de ces ACA est handicapée par les capteurs et actuateurs dont ils disposent. S'ils ouvrent le champ de nos expériences sensorielles à un monde virtuel, l'espace d'interaction est le plus souvent limité à des substituts moteurs et sensoriels appauvris (souris, écran d'ordinateur, etc.) Du fait même de l'appauvrissement de cet espace d'interaction et de leur manque d'emprise physique sur l'environnement, ces créatures nous poussent à renforcer l'effort de compréhension et de modélisation de la structure des signaux de communication résiduels dont ils disposent (parole et gestes).

Cette essentielle mise en présence, cette incarnation du système d'information est souvent oubliée dans la conception d'interfaces personnes-systèmes de plus en plus éclatés dans l'environnement. Le concept « d'intelligence ambiante » parfois dénommé outre-Atlantique « *calcul ubiquitaire* » ou « *ordinateur évanescent* », présente ainsi la vision une fusion progressive des systèmes informationnels et du monde physique. Il préfigure un monde où pratiquement chaque objet de l'environnement, accessible à notre expérience sensible sera équipé de capteurs, de capacité de communication voire de calcul embarqué permettant de le rendre sensibles à nos gestes (par exemple le regard, comme dans « Minority Report ») ou d'exploiter de manière intuitive ses « affordances » naturelles (par exemple exploiter les empreintes digitales de l'utilisateur lors de la préhension). L'environnement réel est en passe de devenir l'interface d'une informatique le plus souvent invisible et ubiquitaire (Van de Velde 2003; Riva, Vatalaro et al. 2005). Cette dématérialisation des interfaces et cette mise en œuvre de technologies attentives capables de prédire nos désirs et nos intentions courent le risque d'être perçues comme intrusives et inquiétantes. Elles posent aussi le problème de la gestion

des erreurs et de l'adaptabilité. Ange-gardien ou génie virtuel phagocytant les objets communicants appropriés au voisinage de l'utilisateur, l'ACA peut ainsi incarner une présence (Riva, Davide et al. 2003) tangible, cohérente et crédible du système d'information, contrastant avec cette intelligence ubiquitaire, éclatée et désincarnée que notre sens commun a du mal à appréhender.

Notre propos n'est pas centré sur les nécessaires compétences linguistiques, paralinguistiques, dialogiques, pragmatiques et situationnelles que de tels agents doivent acquérir pour interpréter nos actions et intentions. Comme nous l'avons souligné plus haut, cette intelligence artificielle doit être incarnée et accessible à l'expérience. Cette intelligence doit réguler des boucles sensori-motrices de plus bas niveau afin que les actions de nos agents, les signaux qu'ils émettent et perçoivent de l'environnement en réaction à leurs actions, maintiennent des représentations cohérentes et intelligibles.

Cet article présente les efforts récents entrepris au sein de l'équipe « Machines Parlantes, Agents Communicants et Interaction Face-à-face » du département « Parole et Cognition » du laboratoire GIPSA, afin de doter un agent conversationnel – incarné sous forme de tête parlante virtuelle – de capacités à interagir en face-à-face avec un partenaire humain. Après une revue rapide des technologies disponibles permettant de bâtir des têtes parlantes, nous détaillons ci-après quelques composantes du système de vision qu'un agent virtuel doit posséder pour signaler sa sensibilité à l'environnement tant du point de vue de sa capacité à analyser une scène visuelle qu'à reproduire les saccades oculaires qu'un humain aurait produites pour y parvenir et aux mouvements des yeux et des paupières qui les auraient accompagnées.

Genèse des têtes parlantes virtuelles

Animation faciale

Les premières techniques d'animation de visages par ordinateur (pour une revue synthétique, voir Parke and Waters 1996) ont conçu l'image d'un visage animé comme la projection par un modèle de caméra d'une structure 3D sous-jacente. Ce maillage 3D est soit passif (il n'exerce aucune résistance à la déformation) avec dans ce cas un contrôle explicite du mouvement de chaque nœud, soit actif (il possède des propriétés mécaniques) avec des forces qui s'appliquent directement sur les nœuds dans le cas de matelas de ressorts (Terzopoulos and Waters 1990) ou le long de fibres musculaires organisées dans l'épaisseur du maillage (Chabanas and Payan 2000). Ces techniques d'animation restent abondamment employées, notamment dans la production de films de synthèse, même si le contrôle de tels modèles est parfois problématique. L'animation de tels modèles à partir de mouvements capturés sur un locuteur pose le problème de mise à l'échelle du modèle 3D (morphologie, déformations intrinsèques) aux caractéristiques anatomiques du sujet. Si le maillage est actif, il reste de plus à estimer les forces à exercer en fonction des déformations observées (Pitermann and Munhall 2001).

Plus récemment, de nouvelles techniques d'animation basée image ont été proposées. Cette classe de techniques cherche à générer des images de synthèse en modifiant des images réelles piochées dans des vidéos : si les premiers systèmes procèdent par interpolation d'images-clés (Ezzat and Poggio 1998 utilisent le flux optique pour interpoler entre visèmes²), les techniques les plus récentes exploitent des modélisations plus complètes, intégrant des modèles de contrôle appris sur les données : leur avantage est d'exploiter les informations de mouvement, de variations de forme et d'apparence contenues

² Un visème est une configuration générique du visage permettant de décrire l'articulation d'un son en contexte. Les sons [u] et [y] correspondent ainsi au même « visème » avec la bouche fermée, arrondie et protruse. Le parallèle avec la notion de phonème – unité phonologique minimale – est abusif.

dans de grandes quantités de vues d'un même visage en mouvement. Des techniques d'apprentissage statistiques permettent d'exploiter de manière optimale ces grandes quantités de données : chaînes de Markov cachées (Tamura, Masuko et al. 1998) pour le contrôle de l'articulation en fonction de la chaîne phonétique à prononcer ou modèles actifs de forme et d'apparence (Cootes, Edwards et al. 2001) pour le rendu de cette articulation.

On voit donc que quelle que soit la technique utilisée (cf. Figure 1), le réalisme des divers modèles (contrôle, forme, apparence) est fortement dépendant de notre capacité à capturer et modéliser les caractéristiques des mouvements biologiques planifiés par la cognition humaine, de même que leurs conséquences visibles et audibles.

Ce lien intime entre signaux et contenu ne s'arrête pas là : les actes de discours et les diverses expressions faciales doivent être correctement décodées par l'interlocuteur humain de ces têtes parlantes. Parfois satisfaisants d'un strict point de vue algorithmique, esthétique ou objectif (les erreurs de reconstruction sont souvent en dessous du millimètre), ces modèles peuvent se révéler décevants en matière de transmission d'informations.

Figure 1 : Têtes parlantes virtuelles. De gauche à droite : Baldi (Massaro 2006); Greta (Poggi, Pelachaud et al. 2005) ; Mother (Revéret, Bailly et al. 2000; Bailly, Elisei et al. 2006); Mary101 (Ezzat, Geiger et al. 2002).

Evaluation

Nous disposons de divers bancs de tests perceptifs exploitant les diverses propriétés de la transmission d'informations phonétiques

par le canal visuel. La vision du visage améliore l'intelligibilité du message sonore (évalué en termes de phonèmes ou mots correctement reconnus par un panel de sujets) en présence de bruit. Le gain d'intelligibilité équivaut à une augmentation du rapport signal sur bruit de plus de 11dB (Summerfield 1991; Benoît and Le Goff 1998). Le canal visuel facilite aussi la compréhension de messages à fort contenu sémantique ou énoncés dans une langue mal maîtrisée par l'auditeur (Reisberg, McLean et al. 1987). La fusion de stimuli audiovisuels incohérents (son et animations faciales non produits par les mêmes mouvements... comme c'est le cas en post-synchronisation) peut générer des percepts de composition : c'est le cas de l'effet McGurk (McGurk and MacDonald 1976), où le son [ba] postsynchronisé avec l'articulation du son [ga] conduit à la perception nette du son [da] ou [va].

Il existe peu de campagnes d'évaluation où la capacité des systèmes d'animation à produire des gestes intelligibles, faciles à comprendre et cohérents avec le son, ait été véritablement testée. On peut bien sûr argumenter que le nombre d'entrées en salle de cinéma ou le nombre de téléspectateurs sont un facteur économique déterminant et suffisant pour la majorité des applications. Pandzic *et al.* (1999) ont comparé l'intelligibilité, la compréhension et l'agrément de trois systèmes d'animation faciale dans une simple tâche de lecture de textes évaluée par un panel de 130 téléspectateurs. Un résultat intéressant ressortant de cette étude – outre que l'animation doit être réalisée à plus de 10 images/seconde – est que la faible intelligibilité d'un système peut être compensée par une mobilisation plus importante des ressources cognitives du spectateur : les bonnes performances en intelligibilité du système jugé le moins agréable (bien que technologiquement le plus abouti) étaient obtenues avec un temps de jugement deux fois supérieur au système jugé le plus agréable.

Après avoir conduit avec succès pour la première fois un test de Turing³ avec leur système Mary, Ezzat et al. (2002) réfèrent notre

³ Dans Ezzat et al (2003), les sujets devaient décider si la vidéo qu'ils voyaient était réelle ou synthétique. Le son original, **par contre**, était conservé.

enthousiasme avec des résultats en intelligibilité assez décevants (Geiger, Ezzat et al. 2003). De même, alors que Baldi, la tête parlante créée par l'équipe de Dominic Massaro à l'USC, est utilisée dans des applications d'enseignement des langues étrangères (Massaro 2006), elle n'arrive pas à reproduire l'effet Mc Gurk (Massaro 1998). Cette congruence, ni nécessaire ni suffisante, entre agrément, intelligibilité et naturel a été confirmée dans d'autres études. Ainsi, si les nouvelles techniques de traitement du signal et des images, les nouveaux outils des mathématiques appliquées et les capacités de traitement et de stockage de grandes masses de données permettent d'atteindre des qualités de synthèse de scènes multimodales d'un réalisme convaincant, il ne faut pas oublier que notre cerveau est difficile à berner : si nous admirons le mime dans sa capacité à imiter les gestes biologiques, nous ne sommes en fait pas trompés sur les causes physiques réelles de ceux-ci (Runeson et Frykholm 1981). Une explication possible – mais dont il **ne** faut pas abuser – réside dans le couplage fort entre perception et action : de plus en plus de données neuroanatomiques (voir notamment la découverte des fameux neurones miroirs par Rizzolatti, Fogassi et al. 2001) montrent que les zones motrices peuvent être recrutées pour analyser l'action de l'autre, et vice-versa que les zones perceptives sont parfois recrutées lors de la programmation ou du suivi en ligne de l'action. La réutilisation de gestes biologiques capturés – souvent utilisée en animation – permet de résoudre une partie du problème, mais cette option ne peut concerner que des situations sans réelle interactivité (diffusion) ou des situations où des schémas pré-établis peuvent faire illusion.

Interaction

Or notre comportement cognitif – et particulièrement langagier – est fortement conditionné par les actions de nos interlocuteurs et de manière plus générale par l'environnement dans lequel l'interaction a lieu. La gestion de nos tours de parole, l'intonation de notre voix, notre style d'élocution, nos expressions faciales sont conditionnés non seulement par notre propre état mental et nos intentions de

communication, mais aussi par la perception immédiate d'évènements surgissant dans l'environnement, par le contenu linguistique et paralinguistique du discours de nos interlocuteurs ainsi que par l'évaluation permanente de leurs caractéristiques individuelles (origine et compétence linguistiques, âge, sexe, position sociale, état émotionnel, etc.) que l'interaction nous permet de mettre à jour. Cette évaluation est évidemment réciproque et c'est un véritable espace de croyance mutuel que nos faits et gestes actualisent (Clark et Brennan 1991). Il semble donc essentiel de comprendre l'influence de ces divers facteurs endogènes et exogènes sur notre comportement.

Nous allons illustrer notre propos par un exemple concret sur lequel notre équipe a porté ses efforts depuis trois ans : la gestion du regard des ACA. Afin d'apporter un semblant de vie au regard souvent vide et désarmé des agents conversationnels, des solutions *ad hoc* (clignements des paupières pseudo-périodiques, mouvements légers de la tête, des sourcils ou du regard préstockés, etc.) sont souvent adoptées. Or si les mouvements de la tête, des sourcils ou du regard sont notamment liés à la structure du discours et des tours de parole, même la fréquence de nos clignements dépend de notre état cognitif : nous avons notamment déterminé que cette fréquence était beaucoup plus importante lorsque nous parlons que lorsque nous écoutons... attentivement !

Regard : phénoménologie et modélisation

Données psychophysiques

La rétine est constituée de deux types de détecteurs : les bâtonnets, cellules sensibles aux changements d'illumination et les cônes, cellules concentrées sur la fovéa, zone centrale permettant l'analyse colorimétrique fine de la scène visuelle. Ces deux types de cellules alimentent majoritairement deux circuits d'analyse visuelle (contraste psychologique ou fonctionnel selon Mishkin, Ungerleider et al. 1983;

Jeannerod 1997). Le circuit dorsal (supposé répondre à la question « Où » et engageant une analyse pragmatique du champ visuel), souvent surnommé détecteur de mouche, est dédié au contrôle de la motricité des yeux. Sensible au mouvement dans la zone périphérique de la rétine, il permet d'amener une zone d'intérêt sur la fovéa et ainsi au circuit ventral (supposé répondre à la question « Quoi » et engageant une analyse plus sémantique du champ visuel) d'analyser plus finement le contenu de cette zone. La trajectoire de notre regard est ainsi constituée de séries de saccades, fixations et poursuites. Nous déclenchons environ 150,000 saccades par jour, transitions très rapides (30 à 120 millisecondes pour des angles de 1 à 40 degrés) entre fixations d'une durée moyenne comprise entre 200 et 300 millisecondes. Ces fixations sont caractérisées par une **convergence** binoculaire et la présence de microsaccades permettant à l'œil de compenser l'adaptation rétinienne. Les saccades sont le résultat d'une analyse perceptive combinant une réactivité à des stimuli exogènes (faits saillants de la scène audiovisuelle) et stimuli endogènes (faits recherchés par une exploration active de la scène). Yarbus (1967) a ainsi montré la perméabilité des stratégies de scrutation de scènes à la tâche cognitive imposée au sujet : saillance et pertinence des objets de la scène proposée participent de manière compétitive à attirer le regard.

Les yeux sont perçus comme des stimuli exogènes particuliers : le regard de l'autre bénéficie d'un traitement particulier par notre système de perception visuel (Driver, Davis et al. 1999). Si nous sommes comme de nombreuses espèces dotés d'un système de vision très sensibles aux regards dirigés vers nous, nous sommes en outre très sensibles aux regards portés sur des objets d'intérêt situés dans notre champ de vision, voire hors de notre champ de vision. Pourtois *et al.* (2004) ont ainsi montré que cette analyse peut conditionner l'interprétation d'expressions faciales. Le système de détection du regard de l'autre est d'ailleurs l'un des modules de base de la plupart des modèles de théorie de l'esprit (Baron-Cohen 1995; Povinelli et Preuss 1995), composante essentielle de notre système cognitif et élément fondateur du développement de notre compréhension du monde,

permettant de prêter et d'interpréter les intentions des autres agents cognitifs présents dans notre environnement immédiat.

Le regard de l'autre biaise nos réactions à des stimuli exogènes : Langton et al (2000) ont ainsi montré qu'un regard concomitant à l'apparition d'un stimulus exogène réduit notre temps de réaction à l'apparition de ce stimulus de manière significative par rapport à un regard non concomitant. Ce bénéfice n'est pas observé si le regard est remplacé par une flèche. Nous avons montré (Raidt, Bailly *et al.* 2006) que le modeste bénéfice obtenu (20 millisecondes) par les stimuli subliminaux utilisés par Langton (affichage d'un visage statique pendant 50 millisecondes) devient substantiel lorsque ce pointage est réalisé par un ACA (200 millisecondes), et plus particulièrement lorsque cette deixis est multimodale (400 millisecondes).

Modèles de contrôle du regard

Deux types de modèles de contrôle du regard ont été proposés dans la littérature : une première classe de modèles s'est particulièrement intéressée à la scrutation de scènes naturelles statiques ou dynamiques (Itti, Dhavale et al. 2003; Sun 2003) en s'attachant à reproduire nos stratégies d'analyse de scènes ; une autre classe de modèles s'est attachée aux propriétés statistiques des saccades oculaires et leur relation aux états cognitifs du locuteur (Lee 2002; Bilvi and Pelachaud 2003). On retrouve ici de manière caricaturale la dichotomie entre réaction aux stimuli exogènes et scrutation endogène. Les situations réelles d'interaction nécessitent bien sûr le recrutement et la collaboration de ces deux mécanismes d'analyse ascendante (réactive à l'environnement) et descendante (proactive et sélectionnant les stimuli en adéquation avec la tâche) : ainsi le modèle proposé par Itti *et al.* (2003) couple un système d'analyse de scène comprenant une carte de saillance, une carte de pertinence et une carte d'attention, avec un modèle biologique de saccades oculaires pour contrôler le regard d'un ACA. En ajoutant un mécanisme de poursuite, la gestion d'une pile d'attention et un système de gestion de la carte d'attention utilisant des modules spécialisés de reconnaissance d'objets (incluant

notamment la reconnaissance de visages), nous avons montré que la trajectoire du regard synthétisée pouvait être très proche de trajectoires oculométriques capturées sur des sujets observant les mêmes scènes visuelles (Picot, Bailly et al. 2006). On note cependant que, pour les observateurs humains, les visages restent beaucoup plus prégnants dans la scène que maints éléments saillants des sujets (gestes, éléments vestimentaires saillants, etc.) ou de l'environnement (saturation ou variation de la lumière, etc.) : la présence de visages dans notre champ de vision capte de manière évidente notre attention. Bouche et yeux sont des éléments susceptibles de délivrer des informations cruciales pour la compréhension de cette scène et il n'est pas surprenant que ces éléments soient scrutés de manière privilégiée lors de visualisations de vidéos de visages parlants (Vatikiotis-Bateson, Eigsti *et al.* 1998).

Il reste que peu de données quantitatives existent sur nos stratégies de scrutation de visages en situation d'interaction face-à-face. Gullberg et al (2001) ont pourtant montré qu'elles étaient significativement différentes des situations artificielles de « voyeurisme » vidéo où l'impact des règles élémentaires de l'interaction sociale est inexistant : si les contacts visuels y sont plus rares et écourtés, les interactions effectives permettent en outre de mesurer l'impact des activités cognitives (écouter, parler, se préparer à parler, réfléchir, signaler que l'on veut prendre la parole ou qu'on est prêt à la céder, etc.) sur nos actions ainsi que les actions de l'autre sur les nôtres. L'attention mutuelle ne peut se mesurer qu'en interaction réelle.

Figure 2 : Conversation face-à-face médiatisée. Le dispositif expérimental (© photothèque CNRS) constitué de deux caméras et écrans couplés permet d'enregistrer de manière synchrone non seulement les signaux audio et vidéo des deux interlocuteurs mais aussi leurs jeux de regard. Les écrans sont en fait équipés d'occulomètres non invasifs (© Tobii) opérant par analyse d'images infrarouges. Les caméras et lumières infrarouges sont intégrées aux écrans. Les traces multimodales collectées lors de tâches finalisées font l'objet d'analyses subséquentes.

Figure 3 : Histogrammes présentant la proportion de saccades oculaires portées par notre locutrice de référence sur diverses zones d'intérêt de son champ de vision (en ordonnée de haut en bas : visage de l'interlocuteur hors bouche et yeux, œil droit/gauche, bouche, hors visage, hors écran) en fonction de son état cognitif (en abscisse de gauche à droite : parler, écouter, attendre, lire, se préparer à parler, réfléchir, autre) lors de conversation face-à-face médiatisées. L'histogramme de gauche présente le profil moyen des échanges où la locutrice conduit l'échange, celui de droite lorsqu'elle répond aux questions. On peut noter l'impact à la fois des états cognitifs et du rôle du locuteur dans la conversation sur ces distributions. A noter, par exemple, la grande proportion de fixations sur la bouche quand elle écoute par rapport à lorsqu'elle parle ainsi que la systématique fixation de l'œil droit de son interlocuteur quand elle se prépare à répondre.

Figure 4. Fréquence de clignement de notre locutrice de référence en fonction de quatre états cognitifs (en abscisse : écouter, attendre, parler, se préparer à parler) lors de son interaction avec 9 interlocuteurs (en ordonnée) quand elle conduit l'échange (à gauche) vs quand elle répond aux questions (à droite). Notez dans les deux cas, la quasi absence de clignements lors de l'écoute et le clignement systématique accompagnant l'initiation de la phonation.

Nouveaux dispositifs expérimentaux

L'étude d'interactions effectives entre humains puis entre ACA et humains doit donc être menée afin de pouvoir disposer de données concernant la boucle de perception/action et sur l'impact de la qualité de l'incarnation de l'information sur la compréhension de cette information, ainsi que les ressources cognitives mobilisées pour gérer cet échange d'informations.

De nouveaux dispositifs d'interaction multimodale doivent ainsi être développés afin de recueillir, simuler, modéliser puis synthétiser en temps réel les signaux de telles interactions. C'est pour financer de tels lieux d'expérimentation que le GIS PEGASUS et le PPF Interactions Multimodales ont été créés sur le site grenoblois, ainsi que le projet Présence financé par le cluster ISLE de la région Rhône-Alpes. La plate-forme MICAL de notre équipe fait partie de ce réseau

de lieux d'expérimentation. Elle nous a permis récemment (Bailly *et al.*, 2010) d'étudier les jeux de regard lors d'interactions face-à-face médiatisées par un système de visiophonie original consistant en deux jeux de caméras et d'oculomètres couplés (voir Figure 2). Cette plateforme nous permet d'enregistrer de manière non invasive les regards, les signaux audio et vidéo reçus par chaque interlocuteur, ainsi que de comparer les jeux de regard en interaction médiatisée avec ceux issus de la scrutation d'une interaction factice (vidéo préenregistrée) ou ceux instanciés par un ACA.

L'analyse de jeux de regard lors de scénarios d'interaction nécessitant une forte attention mutuelle – lecture et répétition de phrases sémantiquement imprédictibles (Benoît, Grice *et al.* 1996) – montre notamment que si l'état cognitif des l'interlocuteurs a un impact sur la distribution des saccades oculaires sur le visage de l'autre, le rôle de chaque partenaire dans la conversation influe sur son comportement (voir Figure 3) : ainsi, le lecteur regarde plus la bouche de son interlocuteur lorsque celui-ci donne les instructions que lorsqu'il doit confirmer qu'il les a bien comprises. De même, nous évitons de cligner des yeux lorsque nous écoutons attentivement notre interlocuteur alors que nous cillons environ une fois toutes les deux secondes lorsque nous parlons (voir Figure 4). Le scénario décrit ci-dessus permet en outre d'étudier l'impact de la perception du regard de l'interlocuteur sur l'autre et ainsi d'intégrer cette composante essentielle de la perception de la scène dans la planification de l'action.

Ces différents niveaux de régulation (réactive, linguistique, paralinguistique, sociale, etc.) de la boucle de perception-action (Thórisson 2002) doivent donc être étudiés en action. Ces plates-formes doivent permettre de confronter de larges panels d'utilisateurs à des boucles d'interaction médiatisées voire entièrement contrôlées par des systèmes artificiels et ceci dans de vraies conditions d'usage, c'est-à-dire opérant en temps réel et avec des sujets naïfs.

Conclusions

L'informatique, les outils et techniques du traitement de signal et des images, de l'automatique, de la modélisation statistique et plus généralement des mathématiques appliquées permettent d'analyser et de capturer les régularités des manifestations observables d'activités cognitives complexes et de progresser ainsi dans notre compréhension des facteurs conditionnant notre comportement. L'identification de ces facteurs permet de concevoir des systèmes d'information plus adaptés à la manière dont notre cerveau traite et assimile ces informations. Une partie essentielle de ce travail d'analyse réside dans le travail préliminaire de conditionnement, de scénarisation des activités cognitives étudiées, de recueil de données sur le comportement observable, de l'identification des circuits de traitement de ces informations et des liens possibles entre signaux observables et activités cognitives supposées ainsi que la nécessaire confrontation des modèles développés aux résultats expérimentaux. Comme nous l'avons vu, les résultats expérimentaux sont conditionnés par les dimensions sociales de nos échanges et de notre perception du monde. Ces dimensions doivent être prises en compte pour l'interprétation des résultats. L'analyse d'un signal aussi simple qu'une trajectoire de regard fait ainsi appel à tout l'éventail des sciences cognitives.

Plus que jamais la cohabitation entre disciplines est nécessaire au sein de nos laboratoires dédiés aux sciences et technologies de l'information. Cet équilibre est difficile à trouver entre l'excellence disciplinaire et la nécessaire prise de risque aux frontières des disciplines. Les sciences cognitives sont ainsi une création originale : loin de constituer une simple discipline d'interface, elles sollicitent de multiples disciplines pour éclairer un système complexe : l'intelligence et ses **gages sensibles**.

Remerciements

Ce travail a été en partie financé par le projet Présence du cluster ISLE de la région Rhône-Alpes (<http://www.grenoble-universites.fr/isle/>), le PPF «Interactions Multimodales» (http://www.icp.inpg.fr/PEGASUS/PPF_IM.html) financé par les quatre universités de Grenoble et le projet ANR « Amorces ».

Références

- Bailly, G., F. Elisei, S. Raidt, A. Casari and A. Picot (2006). Embodied conversational agents : computing and rendering realistic gaze patterns. Pacific Rim Conference on Multimedia Processing, Hangzhou, 9-18.
- Bailly, G., S. Raidt, et al. (2010). "Gaze, conversational agents and face-to-face communication." Speech Communication - special issue on Speech and Face-to-Face Communication **52**(3): 598–612.
- Baron-Cohen, S. (1995). Mindblindness. Boston, MA, MIT Press.
- Benoît, C., M. Grice and V. Hazan (1996). "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences." Speech Communication **18**: 381-392.
- Benoît, C. and B. Le Goff (1998). "Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP." Speech Communication **26**: 117-129.
- Bilvi, M. and C. Pelachaud (2003). Communicative and statistical eye gaze predictions. International conference on Autonomous Agents and Multi-Agent Systems (AAMAS), Melbourne, Australia
- Cassell, J., J. Sullivan, S. Prevost and E. Churchill (2000). Embodied Conversational Agents. Cambridge, MIT Press.
- Chabanas, M. and Y. Payan (2000). A 3D Finite Element model of the face for simulation in plastic and maxillo-facial surgery. International Conference on Medical Image Computing and Computer-Assisted Interventions, Pittsburgh, USA, 1068-1075.
- Clark, H. H. and S. E. Brennan (1991). Grounding in communication. Perspectives in socially shared cognition. L. B. Resnick, J. M. Levine and S. D. Teasley. Washington, DC, American Psychological Association: 127-150.

- Cootes, T. F., G. J. Edwards and C. J. Taylor (2001). "Active Appearance Models." IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(6): 681-685.
- Driver, J., G. Davis, P. Riccardelli, P. Kidd, E. Maxwell and S. Baron-Cohen (1999). "Shared attention and the social brain : gaze perception triggers automatic visuospatial orienting in adults." Visual Cognition **6**(5): 509-540.
- Ezzat, T., G. Geiger and T. Poggio (2002). Trainable videorealistic speech animation. Proceedings of ACM SIGGRAPH, San Antonio, Texas
- Ezzat, T. and T. Poggio (1998). MikeTalk: a talking facial display based on morphing visemes. Computer Animation, Philadelphia, PA, 96-102.
- Geiger, G., T. Ezzat and T. Poggio (2003). Perceptual evaluation of video-realistic speech. Cambridge, MA, Massachusetts Institute of Technology: #224/AI Memo #2003-003.
- Gullberg, M. and K. Holmqvist (2001). Visual attention towards gestures in face-to-face interaction vs on screen. International Gesture Workshop, London, UK, 206-214.
- Itti, L., N. Dhavale and F. Pighin (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, 64-78.
- Jeannerod, M. (1997). The cognitive neuroscience of action. Oxford, UK, Blackwell.
- Kaplan, F. (2006). Les Machines apprivoisées : Comprendre les robots de loisir. Paris, France, Vuibert.
- Langton, S., J. Watt and V. Bruce (2000). "Do the eyes have it ? Cues to the direction of social attention." Trends in Cognitive Sciences **4**(2): 50-59.
- Lee, S. P. (2002). Facial animation system with realistic eye movement based on a cognitive model for virtual agents. Center for Human Modeling and Simulation. Philadelphia, PA, University of Pennsylvania: 100.
- Massaro, D. (1998). Illusions and issues in bimodal speech perception. Auditory-Visual Speech Processing Conference, Terrigal, Sydney, Australia, 21-26.
- Massaro, D. W. (2006). A computer-animated tutor for language learning: Research and applications. Advances in the spoken language development of deaf and hard-of-hearing children. P. E. Spencer and M. Marshark. New York, NY, Oxford University Press: 212-243.

- McGurk, H. and J. MacDonald (1976). "Hearing lips and seeing voices." Nature **264**: 746-748.
- Mishkin, M., L. G. Ungerleider and K. A. Macko (1983). "Object vision and spatial vision: two cortical pathways." Trends in Neuroscience **6**: 414-417.
- Pandzic, I., J. Ostermann and D. Millen (1999). "Users evaluation: synthetic talking faces for interactive services." The Visual Computer **15**: 330-340.
- Parke, F. I. and K. Waters (1996). Computer Facial Animation. Wellesley, MA, USA, A.K. Peters.
- Picot, A., G. Bailly, F. Elisei and S. Raidt (2006). Scrutation de scènes naturelles par un agent conversationnel animé. Workshop sur les Agents Conversationnels Animés, Toulouse, France
- Piternann, M. and K. G. Munhall (2001). "An inverse dynamics approach to facial animation." Journal of the Acoustical Society of America **110**: 1570-1580.
- Poggi, I., C. Pelachaud, F. de Rosi, V. Carofiglio and B. de Carolis (2005). GRETA. A believable embodied conversational agent. Multimodal intelligent information presentation. O. Stock and M. Zancarno. Dordrecht, Kluwer: 3-26.
- Pourtois, G., D. Sander, M. Andres, D. Grandjean, L. Révész, E. Olivier and P. Vuilleumier (2004). "Dissociable roles of the human somatosensory and superior temporal cortices for processing social face signals." European Journal of Neuroscience **20**: 3507-3515.
- Povinelli, D. J. and T. M. Preuss (1995). "Theory of Mind: evolutionary history of a cognitive specialization." Trends in Neuroscience **18**(9): 418-424.
- Raidt, S., G. Bailly and F. Elisei (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, 2544-2549.
- Reisberg, D., J. McLean and A. Goldfield (1987). Easy to hear but hard to understand: a lipreading advantage with intact auditory stimuli. Hearing by Eye: The Psychology of LipReading. B. Dodd and R. Campbell. Hillsdale, New Jersey, Lawrence Erlbaum Associates: 97-113.
- Revéret, L., G. Bailly and P. Badin (2000). MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. International Conference on Speech and Language Processing, Beijing, China, 755-758.

- Riva, G., F. Davide and W. A. Ijsselstein (2003). Being there: concepts, effects and measurements of user presence in synthetic environments. Amsterdam, IOS Press.
- Riva, G., F. Vatalaro, F. Davide and M. Alcañiz (2005). Ambient Intelligence. The evolution of technology, communication and cognition towards the future of human-computer interaction. Amsterdam, OCSL Press.
- Rizzolatti G., Fogassi L., Gallese V. 2001. « Neurophysiological mechanisms underlying the understanding and imitation of action ». *National Review of Neuroscience*, 2: 661-670.
- Runeson, S. and G. Frykholm (1981). "Visual perception of lifted weight." Journal of Experimental Psychology: Human Perception and Performance 7: 733-740.
- Summerfield, Q. (1991). Visual perception of phonetic gestures. Modularity and the motor theory of speech perception. I. G. Mattingly and M. Studdert-Kennedy. Hillsdale, NJ, Lawrence Erlbaum Associates: 117-138.
- 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.
- Tamura, M., T. Masuko, T. Kobayashi and K. Tokuda (1998). Visual speech synthesis based on parameter generation from HMM: speech-driven and text-and-speech-driven approaches. Auditory-visual Speech Processing Workshop, Terrigal, Sydney, Australia, 219-224.
- Terzopoulos, D. and K. Waters (1990). "Physically-based facial modeling, analysis and animation." The Journal of Visual and Computer Animation 1: 73-80.
- 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.
- Van de Velde, W. (2003). The world as computer. Smart Objects Conference, Grenoble, France, 27-29.
- Vatikiotis-Bateson, E., I.-M. Eigsti, S. Yano and K. G. Munhall (1998). "Eye movement of perceivers during audiovisual speech perception." Perception & Psychophysics 60: 926-940.
- Yarbus, A. L. (1967). Eye movements during perception of complex objects. Eye Movements and Vision. L. A. Riggs. New York, Plenum Press. VII: 171-196.