

Rapport d'activités 2002 - 2006

Gérard BAILLY, DR2 CNRS

<http://www.icp.inpg.fr/~bailly>

Institut de la Communication Parlée UMR5009

INPG – Université Stendhal, 46, av. Félix Viallet – 38031 GRENOBLE

Domaines de recherche

Mes activités de recherche de base portent sur l'étude **des signaux de la communication parlée et la conception et l'évaluation d'agents conversationnels animés** capables d'engager une interaction face-à-face multimodale avec un usager. Ces recherches s'appuient sur des **plate-formes expérimentales** permettant de recueillir les signaux multimodaux dans des situations contrôlées, lors d'interactions humaines face-à-face ou d'interaction d'utilisateurs avec les systèmes interactifs temps-réel que nous concevons.

1. **Signaux de la communication parlée face-à-face.** La parole est l'un des moyens les plus sophistiqués pour interagir avec nos semblables. Nos actes de langage sollicitent tous les segments du corps, non seulement ceux nous permettant de mettre en forme la parole (mâchoire, larynx, langue, lèvres, etc.) avec des conséquences audibles et visibles mais aussi les mouvements de la tête et des mains, du regard et même du buste. Le regard ainsi participe au maintien de l'attention, à des gestes iconiques ou déictiques ainsi qu'à signaler diverses activités cognitives concomitantes ou préalables à l'acte de langage. Ces signaux/mouvements présentent une importante variabilité et participent – chacun avec ses propres contraintes et structures - à l'encodage de multiples fonctions discursives. Des modèles de synchronisation, de fusion et d'encodage fonctionnel sont nécessaires afin de pouvoir analyser, comprendre et reproduire les signaux multimodaux observés lors d'interactions vraies. Longtemps cantonnées à la lecture de textes, les recherches en synthèse de parole à l'ICP se sont progressivement étendues à des boucles d'interactions plus complexes. La programmation de l'équipe est clairement à présent orientée vers des modèles intégrant analyse et synthèse de scènes multimodales.
2. **Agents conversationnels et interaction multimodale.** L'équipe Machines Parlantes est connue pour ses clones animés (Odisio and Bailly 2004; Bailly, Béjar et al. 2003) et pour son projet de développement d'atlas articulé de la tête nous avons notamment produit le premier modèle tridimensionnel de langue basé sur des données (Badin, Bailly et al. 2002). Notre ambition est de mettre ces têtes parlantes en action dans des scénarios réalistes d'interaction multimodale avec des interlocuteurs humains. La recherche sur les agents conversationnels animés (voir le livre *Principes* de Cassell, Sullivan et al. 2000) largement inspirée par les robots sociaux (Brooks, Breazeal et al. 1999; Brooks 2001) est en pleine mutation. L'agent virtuel comme le robot doit prendre en compte les (ré)actions de son interlocuteur et de l'environnement d'interaction afin d'assurer la pertinence de ses propres actions et des informations qu'il délivre à son partenaire de communication. Médiateur entre l'utilisateur et le système d'information, l'agent conversationnel doit donc posséder un système sophistiqué d'analyse et de compréhension de la scène multimodale dans laquelle il intervient. Pour un système d'interaction multimodale, analyse et synthèse de scènes multimodales sont donc indissociables : le couplage entre ces champs de recherche ouvre de vastes perspectives d'innovation notamment par le couplage nécessaire entre modèles. Notre équipe a d'ailleurs une longue tradition de conception de systèmes d'analyse par la synthèse ou d'analyse basée modèles.
3. **Plate-formes expérimentales.** Ces recherches ne se conçoivent que si nous avons les moyens de capturer l'ensemble des signaux multimodaux de la communication dans des situations aussi réalistes que possible. Tous les projets dans lesquels notre équipe s'est engagé ont contribué à alimenter dans un premier temps la plate-forme expérimentale de l'ICP-Campus dédiée à la capture invasive de signaux : nous avons ainsi financé l'acquisition d'une plate-forme d'enregistrement multi-caméras. Nous sommes aussi de gros consommateurs de capture de mouvement par vidéo, articulographie, cinéradiographie et IRM anatomique pour lesquels nous avons des collaborations extérieures avec des hôpitaux et studios d'animation. Depuis 2004, nous avons identifié le besoin d'une plate-forme expérimentale spécifique pour recueillir les signaux d'interaction multimodale face-à-face de sujets avec des complices humains voire des systèmes d'interaction temps-réel. Nos systèmes d'interaction ne sont plus – ou pas seulement – des vecteurs de communication destinés à montrer notre savoir-faire mais de véritables objets d'étude, des outils de recherche permettant d'avoir accès à des signaux d'interaction réels. Ces plate-formes – parfois lourdes en matière d'investissement et en maintenance technique - sont aussi nécessaires à la conception de systèmes d'interaction que les plate-formes des micro-nanotechnologies. La salle

MICAL est à présent intégrée au réseaux de plate-formes expérimentales du GIS PEGASUS et un ingénieur de recherches récemment recruté au CNRS y a été affecté.

Bilan scientifique

Avant-propos

Il est difficile de séparer la présentation du bilan scientifique personnel d'un responsable d'équipe de celui de son équipe. J'ai impulsé un certain nombre de recherches et de projets qui ont modifié les enjeux scientifiques et les contours mêmes de l'équipe (notamment lors de la fusion des équipes « Machines Parlantes » et « Traitement du signal et Codage » du quadriennal 2003-2007 dans une seule équipe « Machines parlantes, agents communicants & interaction face-à-face » dont j'assurerai la responsabilité pour le prochain quadriennal). Cette réflexion n'est rendue possible et ne débouche sur l'action que grâce à l'adhésion effective des membres de l'équipe. Je ne pourrais porter certains projets collaboratifs sans la crédibilité nécessaire fournie par une équipe dynamique et compétente.

Cette osmose entre forces de proposition et adhérences des acteurs au projet est nécessaire à la fois à façonner le projet scientifique et à l'assurance du succès de sa programmation. J'ai vécu cette mise en forme à de nombreuses reprises et de nombreuses échelles : à celle de mon équipe, de mon laboratoire (j'ai activement participé à la définition du projet de l'ICP du quadriennal 2003-2007 dans lequel le triptyque « Signal, Langage & Cognition » décrit de manière claire les fondements de notre association), de la fédération de laboratoires ELESA (j'ai activement participé à la définition du projet du quadriennal 2003-2007 d'ELESA dans lequel est apparu la notion de secteur-expert, ELESA étant ainsi positionnée en trois grands secteurs : « Energie », « Micro-systèmes » et « Information, Communication & Cognition » que j'anime) ainsi que du GIS PEGASUS (j'ai activement participé à la définition du projet du GIS et au premier appel à propositions PRESENCE).

Je souligne finalement que J.M. Chassery, directeur du laboratoire GIPSA (TGL qui regroupera les forces des laboratoires ICP, LAG et LIS) m'a demandé d'animer une commission scientifique au sein du laboratoire, qui outre un rôle d'incubateur de projets doté de fonds propres, aura aussi un rôle-clé dans la stratégie scientifique du laboratoire.

Je vais cependant dissocier le niveau de l'équipe des autres, où l'empreinte de la volonté scientifique des animateurs est plus patente notamment de part les incitations financières dont ils disposent.

Personnel

Responsabilités dans des structures de recherche

Cette période coïncide avec de multiples prises de responsabilité au sein de structures locales (notamment ELESA et le GIS PEGASUS) et internationales (notamment le Special Interest Group (SIG) Synthèse de l'ISCA).

ELESA. Comme directeur adjoint d'ELESA et comme co-animateur du secteur expert « Information, Communication et Cognition », j'ai contribué à la l'identification dans le bassin grenoblois d'un pôle de compétence sur les « Objets, Agents & Environnements Communicants ». Avec mes collègues Jean Caelen de l'IMAG et Philippe Mallein (MSH-Alpes), nous avons multiplié les initiatives (Projet inter-fédérations PRESENCE, projets soumis au Pôle Européen...) afin de faire identifier par les tutelles (fédérations, universités) la nécessité de faire émerger cette thématique transversale, complémentaire aux deux « poids lourds » verticaux grenoblois que sont le logiciel et les micro-nanotechnologies.

PEGASUS. Nous poursuivons nos efforts de structuration de cette activité au sein du groupement d'intérêt scientifique (PEGASUS) dont je suis le directeur officiel depuis août 2004. Financé principalement par le Plan Pluri-Formation « Objets, Agents & Environnements Communicants » déposé par les 4 universités, PEGASUS nous permet de financer des plate-formes sur l'étude, la simulation et l'évaluation de systèmes d'interaction humaine et de lancer cet automne un appel d'offres à projets autour de cette thématique. Je reviens plus largement sur ce projet scientifique plus bas.

SIG Speech Synthesis. Avec mon collègue et ami Christian Benoît, nous avons organisé en 1990 le premier séminaire international sur la synthèse de parole à Autrans dans les environs de Grenoble. Une sélection de papiers avaient été édités dans un livre emblématique « Talking Machines : data, models & theories » (Bailly and Benoît 1992). Après New Paltz (USA, 1994), Jenolan (Australie, 1998), Perthshire (Ecosse, 2001), Pittsburgh

(USA, 2004), la 6^{ème} édition aura lieu à Bonn en 2006. Quatre laboratoires se sont proposés pour l'organisation de cette manifestation et, pour la première fois, c'est l'ensemble de 173 membres du SIG Speech Synthesis de l'ISCA (International Speech Communication Association) qui a décidé de cette désignation via la liste de discussion que je gère à l'ICP en tant que trésorier du SIG. De 2002 à 2005, Bernt Moebius de l'IMS de Stuttgart, Nick Campbell des laboratoires ATR de Nara – Japon et moi-même avons impulsé une animation plus participative, tout en maintenant notre travail de réflexion en amont (Bailly, Campbell et al. 2003).

Organisation de conférences internationales

CONGRES PAROLE. Avec mon collègue Pierre Badin, nous avons organisé en 2000 les 23^{èmes} Journées d'Étude sur la Parole (Bailly and Badin 2000). En tant que membre du conseil du SIG Speech Synthesis, j'ai participé à l'organisation du 5^{ème} séminaire international sur la synthèse de parole de Pittsburgh (USA, 2004) où je fus conférencier invité (Bailly 2001 ; édité ensuite dans Bailly, Béjar et al. 2003). La création du laboratoire « Signaux, Systèmes, Parole et Cognition » pour le prochain quadriennal devrait fournir la logistique nécessaire à l'organisation d'un congrès majeur du circuit international (notamment ICSLP).

OBJETS COMMUNICANTS. Comme représentant du CNRS, j'ai co-organisé avec Patrice Senn de France-Télécoms R&D la deuxième édition – à présent internationale – de la Conférence sur les Objets Communicants (sOc'2003) à Grenoble. Je co-organise cette année la troisième édition avec mon collègue James L. Crowley de l'INRIA. sOc-EUSAI'2005 (www.soc-eusai2005.org) est devenue pour l'occasion la conférence sur les Objets Communicants et l'Intelligence Ambiante, rassemblant ainsi la communauté SOC et celle d'EUSAI (European Symposium on Ambient Intelligence), série de conférences initiée par Philips à Eindhoven. Nous avons pérennisé ce rendez-vous bi-annuel grâce à la constitution d'un comité de pilotage permanent présidé par Patrice Senn de France Telecoms R&D et Emile Aarts de Philips. Le prochain rendez-vous est d'ailleurs fixé au printemps 2007 à Héraklion en crête et sera co-organisé par Constantine Stephanidis, directeur du laboratoire d'informatique d'ICS-FORTH et Norbert Streitz, directeur du projet AMBIENTE au Fraunhofer IPSI. Je suis dans le comité d'organisation.

Comités scientifiques de conférences internationales

Je suis sollicité pour participer à de nombreux comités scientifiques de conférences internationales : Pour ne citer que les principales, j'ai participé au comité scientifique d' Handicap (Paris, 2006), ETRW Workshop on Experimental Linguistics (Athènes, 2006), LREC'2006 Workshop on Multimodal corpora (Gênes, 2006), Speech Prosody (Dresden, 2006), sOc-EUSAI'2005 (Grenoble, 2005), 1st International Conference on Affective Computing and Intelligent Interaction (Beijing, 2005), EUSIPCO (Vienne, 2004 et Antalya, 2005). Les organisateurs du 16^{ème} Congrès International de Phonétique à Saarbrücken m'ont demandé de faire partie du comité scientifique international qui aura à charge la gestion des revues de larges secteurs de recherche.

Pilotage de projets

Cette période coïncide aussi avec un fort développement des activités contractuelles de l'équipe ... avec un effectif restreint (trois permanents, trois thésards et un contractuel). L'équipe a ainsi largement contribué au budget non consolidé du laboratoire en enchaînant trois projets RNRT (TempoValse, ARTUS et EvaSy) et un projet ROBEA (HR+) dont je suis le responsable scientifique ainsi que deux projets financés par le CNRS (Cognitive et Jeune équipe) dont Denis Beauteemps est le responsable.

Publications

Je suis auteur de 21 publications dans des journaux internationaux, 14 chapitres de livres et 97 articles dans des conférences internationales (voir productions scientifiques). Coéditeur de deux livres (Bailly and Benoît 1992 ; Keller, Bailly et al. 2002), je travaille avec Eric Vatikiotis-Bateson (UBC Vancouver) et mon collègue Pascal Perrier depuis deux ans sur le projet éditorial du livre « Audiovisual Speech Processing ». La sortie du livre (Vatikiotis-Bateson, Bailly et al. 2005) a été annoncé par MIT Press pour septembre 2005 (avec un volume de 250 pages). Grâce à la large couverture des thématiques et au renom des contributeurs, ce livre devrait être une référence incontournable des travaux en traitement audiovisuel de la parole.

Revues

Outre l'activité régulière de revue d'articles soumis à des journaux (dont Journal of Acoustical Society of America, Journal of Phonetics, Speech Communication, IEEE Transactions on Speech and Audio Processing et Acta Acoustica) et à des workshops/conférences internationales, je suis impliqué dans la revue de projets européens : je suis le projet Européen MULTISENSE (<http://sirio.cineca.it/B3C/multisense/>) depuis 2003 et le réseau d'excellence VISNET (<http://www.visnet-noe.org/>) depuis 2004. Je suis en outre membre du conseil scientifique du CEA/DIST depuis 2005. J'ai été sollicité à trois reprises pour expertiser des projets soumis au conseil régional de Lorraine dans le cadre du Pôle de Recherche Scientifique et Technologique.

Encadrement de la recherche

Chaque année plusieurs étudiants travaillent régulièrement sous ma direction. Depuis mon habilitation à diriger des recherches défendue en 2000 et mon implication dans la direction d'ELESA, la plupart de mes thésards sont co-encadrés par un collègue du laboratoire LIS :

1. Directeur de thèse INPG de S. Raidt (depuis 2004) avec L. Bonnaud (LIS)
Communication face-à-face entre un locuteur réel et un clone parlant. Contact visuel et monstration multimodale dans un univers virtuel.
Thèse de troisième cycle, Spécialité Signal, Image, Parole & Télécommunications, Institut National Polytechnique de Grenoble, Grenoble - France.
1. Directeur de thèse INPG de O. Govokhina (depuis 2004) en partenariat avec France Télécoms R&D Rennes
Etude et modélisation de la coarticulation pour l'animation d'un humanoïde de synthèse
Thèse de troisième cycle, Spécialité Signal, Image, Parole & Télécommunications, Institut National Polytechnique de Grenoble, Grenoble - France.
1. Directeur de thèse INPG de M. Bérrar (depuis 2003) avec M. Desvignes (LIS)
Modèles statistiques de la variabilité d'organes articulés du corps humain.
Thèse de troisième cycle, Spécialité Signal, Image, Parole & Télécommunications, Institut National Polytechnique de Grenoble, Grenoble - France.
1. Directeur de thèse INPG de G. Gibert (depuis 2002)
Synthèse multimodale de la parole
Thèse de troisième cycle, Spécialité Signal, Image, Parole & Télécommunications, Institut National Polytechnique de Grenoble, Grenoble - France.
1. Co-directeur de thèse INPG de P. Gacon (depuis 2002) avec P.Y. Coulon (LIS)
Analyse d'images et modèles de formes pour la détection et la reconnaissance. Application aux visages.
Thèse de troisième cycle, Spécialité Signal, Image, Parole & Télécommunications, Institut National Polytechnique de Grenoble, Grenoble - France.
1. Directeur de thèse INPG de M. Odisio (depuis 2000)
Analyse et synthèse de scènes audiovisuelles
Thèse de troisième cycle, Spécialité Signal, Image, Parole & Télécommunications, Institut National Polytechnique de Grenoble, Grenoble - France.
1. Directeur de thèse INPG de B. Holm (1998-2003)
La prosodie de l'énonciation de formules mathématiques
Thèse de troisième cycle, Spécialité Sciences Cognitives, Institut National Polytechnique de Grenoble, Grenoble – France, 2003.

Enseignement

J'ai une activité régulière d'intervention en cursus de DEA et d'année terminale d'école d'ingénieur. Cette activité fait à mon sens partie intégrante de la mission de la recherche publique et permet d'attirer des talents divers vers la recherche. J'interviens actuellement dans les établissements suivants :

1. Cours de la filière « Images et réalité virtuelle » de 3^{me} année ENSIMAG depuis 2001 « Parole et langage: systèmes d'interaction verbale ». Il est à noter que cette filière créée en 2001 est devenue la deuxième en volume d'étudiants de l'école. Mon cours passera de 10,5 à 18 heures cette année.
1. Cours de DEA Sciences Cognitives depuis 1998 « Synthèse de la parole »

Diffusion de la culture scientifique

Parole, têtes parlantes et objets communicants sont des thèmes qui ont bonne presse. J'ai essayé de maintenir une liste à jour des différentes interventions radiophoniques/télévisuelles et éditoriaux où j'ai présenté/défendu diverses facettes de ma recherche personnelle, celle de mon équipe ou des diverses entités que j'anime. Depuis 2001, je suis intervenu dans les médias suivants :

- **CNRS Info** (2001) *Les clones ont la parole* , Paris, Octobre-Novembre, p. 21-22.
- **L'Hebdo** (2001) *Nos clones virtuels apprennent à parler* , 49, Lausanne - Suisse, Décembre, p. 76-77.
- **Pour La Science** (2002) *Le vrai visage de la parole* , 291, Paris, Janvier, p. 14.
- **Le Monde** (2002) [Des clones virtuels parlants pour mieux comprendre le langage](#) , Paris, Vendredi 1er Mars, p. 23.
- **01 informatique** (2002) [Les chercheurs truffent notre quotidien d'objets communicants](#) , Paris, Vendredi 22 Novembre, p. 30-32.
- **01 net** (2003) [Les applications industrielles des objets communicants vont se multiplier](#) , Paris, Vendredi 23 Mai, 7:00.
- **France Info** (2003) [L'intelligence ambiante](#) , Paris, Chronique du 16 mai.
- **FR3 Rhône-alpes** (2003) [Reportage sur les démos sOc'2003](#) , 16 mai, 19:00-20:00.
- **Europe1** (2003) [Les objets qui communiquent, Chronique d'Alain Cirou \(avec Patrice Senn, FT R&D\)](#) , 24 Août, 13:15-14:00.
- **Ingénieurs INPG** (2003) *Virtuels clones* (avec Lionel Revéret, GRAVIR/IMAG). **03-3**: 31.
- **I-MAG magazine** (2004) *Clones parlants virtuels : de véritables partenaires de communication entre les usagers, le monde physique et le cyber-monde ?* (avec Lionel Revéret, GRAVIR/IMAG) p. 32.
- **Electronique International Hebdo** (2004) "Comment faciliter l'interaction entre l'homme et les machines intelligentes ?", n°548, 19 février, p.29.
- **Ingénieurs INPG** (2004) *Numéro spécial sur les « objets communicants »*. **04-02**. Editorial et 2diteur associé du dossier.

De l'équipe...

L'équipe « Machines Parlantes » que j'anime développe des systèmes de communication multimodaux inspirés du vivant, ce qui engendre une forte synergie entre expérimentation, analyse, modélisation et évaluation – boucle de conception de « machines parlantes » que j'ai déjà décrit dans mon précédent rapport d'activité et qui a fait l'objet du cadrage de nos recherches pour le présent quadriennal. Deux thèmes de recherche se sont particulièrement développés lors de ces deux dernières années :

- **Développement du modèle prosodique SFC** : j'ai encadré trois thèses sur l'élaboration de ce modèle (Barbosa 1991; Morlec 1997 ; Holm 2003). Le système d'apprentissage automatique développé par Bleike Holm dans le cadre de sa thèse (Holm, Bailly et al. 1999; Holm and Bailly 2000a; Holm and Bailly 2000b; Holm and Bailly 2002; Holm 2003) constitue une étape capitale dans l'implémentation d'un modèle de génération de la prosodie compatible avec les propositions théoriques de V. Aubergé (Aubergé and Bailly 1995; Aubergé 1992). Nous avons résolu le problème – initialement mal posé – de l'apprentissage automatique de ce modèle depuis la simple observation de réalisations où se trouvent suffisamment instanciées les fonctions discursives que l'on désire étudier. Cet apprentissage n'est pas hiérarchique mais compétitif : il permet donc de n'avoir aucun a priori sur l'importance relative des fonctions ni sur les formes des contours qui les véhiculent. Lors du stage de Stefan Raidt de l'université de Dresden et en collaboration avec Hansjoerg Mixdorff de l'université de Berlin , nous avons pu tester ce modèle sur l'allemand et comparer les performances du SFC et du modèle MGI sur un corpus bilingue de formules mathématiques oralisées (Raidt, Bailly et al. 2004). Plus récemment, nous avons appliqué le modèle SFC à une langue à tons (i.e. chinois mandarin) dans le cadre d'une collaboration avec Gaopeng Chen de l'université de Chine à Shanghai (Chen, Bailly et al. 2004).

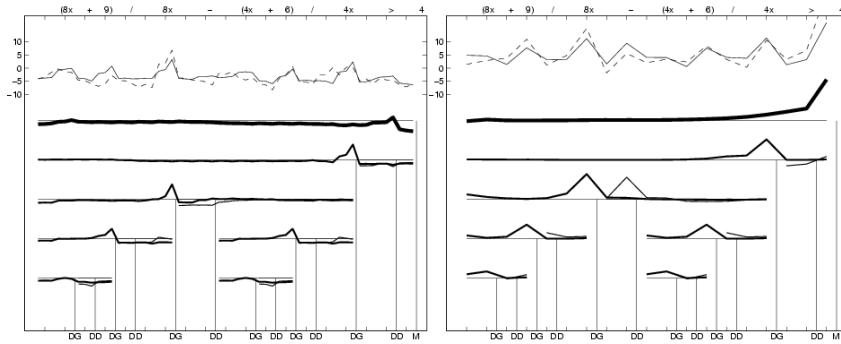


Figure 1 : Résultat de la décomposition automatique de la prosodie d'une formule mathématique énoncée en allemand. A gauche, la mélodie et à droite, la courbe rythmique.

- **Etude, modélisation, suivi et synthèse des gestes faciaux en parole** : grâce à une masse critique de chercheurs (Pierre Badin, Frédéric Elisei de l'équipe, thèses de Mathias Odision, Maxime Bélar et de Pierre Gacon et plus récemment des thèses d' Oxana Govokhina et d'Antoine Serrurier, en collaboration avec Pierre-Yves Coulon et Michel Desvignes du LIS et de Gaspard Breton de France Télécoms R&D), nous avons pris position dans le champ de l'animation faciale (voir notamment les articles acceptés dans des revues tels que Odision and Bailly 2004 ; Bailly, Bélar et al. 2003 ; Badin, Bailly et al. 2002). Grâce à ma collaboration avec Pierre Badin, ce thème de recherches initié et développé au sein de l'équipe par le défunt Christian Benoît (Benoît and Le Goff 1998) s'est fort développé et participe grandement au rayonnement local et international du laboratoire. Les « Machines Parlantes » de l'ICP font partie du cercle restreint des modèles d'articulation faciale capables d'assurer un suivi robuste et une reproduction fine des gestes du visage.

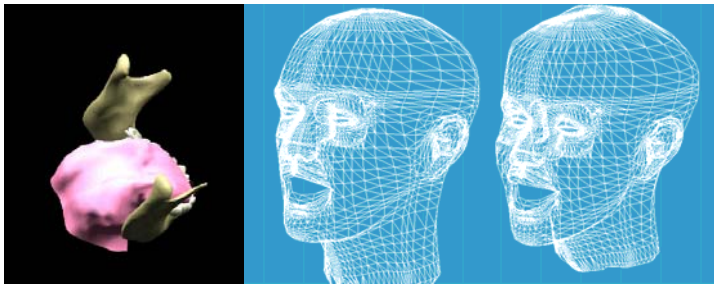


Figure 2 : A gauche le modèle de langue élaboré au sein de l'équipe. A droite : Premier atlas générique articulé paramétrable par les caractéristiques anatomiques du locuteur cible (chaque forme correspond à un geste élémentaire de descente de la mâchoire effectué par deux locuteurs distincts)

- **Etude, modélisation, synthèse et évaluation du Langage Parlé Complété (LPC)** : grâce à de nombreuses sources de financement et une masse critique de chercheurs (Denis Beautemps de l'équipe, thèses de V. Attina et de G. Gibert, en collaboration avec M.A. Cathiard de l'équipe Perception...), nous avons pris position dans le champ de l'étude du LPC, notamment sur l'étude de sa production et de sa synthèse (voir notamment les articles acceptés dans des revues tels que Gibert, Bailly et al. accepted ; Attina, Beautemps et al. 2004). Grâce à un dispositif expérimental entièrement renouvelé par les contrats de l'équipe et notre collaboration dans le projet RNRT ARTUS avec Attitude Studio, nous avons collecté des données de capture de mouvement uniques et précieuses sur le LPC en action. L'étude d'une codeuse oraliste nous permet de bénéficier d'un cadre d'observation que nous maîtrisons bien : le langage articulé et sa dimension audio-visuelle. La structure phonologique particulière du LPC (typiquement Consonne-Voyelle) et le nécessaire contrôle du mouvement de « nouveaux » segments de parole (i.e. le bras, la main et la tête) impose des contraintes de synchronisation avec la production vocale. Cet objet d'étude renouvelle radicalement les enjeux de la parole multimodale et donne de nouveaux thèmes de recherche à l'équipe. Ainsi les mouvements de la tête, traditionnellement recrutés pour véhiculer la structuration discursive ou plus généralement des informations suprasegmentales, sont aussi utilisés pour porter le visage vers la main. Le LPC peut en fait être vu comme la réalisation de constrictions main/visage en des lieux (œil, nez, joue, lèvres, menton...) permettant de désambiguïser la voyelle avec une clé de doigts permettant de désambiguïser la consonne. On retrouve ainsi des thématiques de contrôle moteur, d'équifinalité et de coarticulation, que nous avons déjà abordé en contrôle moteur des articulateurs de la parole. Le LPC constitue aussi un défi intéressant pour le modèle SFC en offrant un nouveau jeu de paramètres caractéristiques à générer (contrôlant les mouvements de la tête) participant largement à l'encodage des structures segmentales et suprasegmentales.



Figure 3 : Capture du mouvement de la main et du visage d'une codeuse LPC. À gauche, Le clone ARTUS animé par un modèle articulé piloté par 9 paramètres pour l'articulation de la main, 7 pour le visage ainsi que 5 paramètres contrôlant les mouvements respectifs de la tête et de la main.

L'étude et la caractérisation du mouvement des segments du LPC (notamment celui de la main) repose aussi de manière cruciale le problème du modèle direct ou comment faire émerger des variables de contrôle proximales (angle des doigts, mouvement du pouce, supination du poignet...) de l'observation du corps en mouvement. La Figure 3 illustre nos premiers résultats de modélisation non-linéaire. Ce problème a fait l'objet du projet « Vésale : Atlas articulé du visage et de la main » financé par le BQR INPG et qui a permis de faire bénéficier Maxime Béar d'une bourse fléchée pour une thèse en co-encadrement entre trois laboratoires grenoblois (ICP, LIS et TIM-C).



Figure 4 : Système de réalité partagée avec interaction multimodale intuitive. L'écran est équipé du système de suivi de regard de la société Tobii. Le clone virtuel au centre de l'écran signale par des saccades oculaires qu'il connaît le centre d'intérêt de l'utilisateur. Dès que le temps de fixation sur un objet (réel ou virtuel) est supérieur à 300ms ou que l'utilisateur dit « ça ! », le clone donne oralement des renseignements sur ce dernier.

Thèmes de recherche prospectifs

Communication multimodale

Les thèmes de recherche introduits plus haut suffisent largement à maintenir une activité de recherche de niveau international. J'ai cependant initié une réflexion au sein de l'équipe, du laboratoire et plus largement au sein du secteur expert d'ELESA sur les termes de l'interaction humaine avec un agent communicant.

En effet, si les thématiques « synthèse de parole expressive » voire « synthèse de parole émotionnelle » deviennent de plus en plus sensibles et mobilisent un nombre croissant de chercheurs, ces thèmes de recherche ne sont souvent entrevues que comme une manière d'accroître la complexité du substrat phonétique et des diverses fonctions de communication impliquées dans une « vraie » interaction. Or cette interaction suppose (et supposait déjà... même lorsqu'on se limite à la lecture de textes) un partenaire de communication... complètement absent de l'échange d'informations. Nos synthétiseurs parlent dans un espace vide et à destination d'inconnus... même si les systèmes de dialogue homme-machine (ou plus récemment personne-système) tente de gérer un espace de représentations communes (historique du dialogue, sémantique des actions...) mais comme si on connectait des entités de traitement de l'information pures et désincarnées...

Un des points de départ de cette réflexion a été nos efforts de mise en scène de clones parlants dans des applications de communication. Dans TempoValse, les usagers testés dans le laboratoire d'usage de France Télécoms R&D ont tous souligné l'importance du contact visuel : ainsi quand la caméra et l'écran d'un terminal de visiophonie ne sont pas dans le même axe de vue, les usagers ont énormément de gêne car l'objet d'attention du regard de l'autre semble hors de leur champ de vision.

L'autre source constante d'inspiration est le travail de l'équipe « Anthropologie Linguistique de la Parole » sur le suivi d'agents et le développement du langage. Il y a apparemment que les systèmes « Où » (Where) et « Qui » (Who) sont mis en place de manière nettement plus précoce que le « Quoi » (What) et que ces systèmes exercent des fonctions largement préemptives sur notre analyse d'une scène multimodale. Avant donc de délivrer une information linguistique, il faut savoir où et à qui cette information doit être délivrée. Cette collaboration a débouché sur un texte commun « Biocybernétique de la communication parlée face à face » que nous avons présenté avec C. Abry au Pôle Européen des Universités de Grenoble. Au système ordonné WWW (Where, Who, What), j'ajouterai le « Comment » (How) cher aux paradigmes du « traitement affectif » (Affective Computing : <http://affect.media.mit.edu/>) et de « robotique incarnée » (Embodied robotics), tous deux mis en projet au MIT : *les signaux support de l'information linguistique sont aussi importants – sinon plus – que l'information elle-même*. En effet, le décodage correct et robuste de l'information nécessite un contrôle fin de multitudes de signaux susceptibles de maintenir l'attention partagée des interlocuteurs sur le discours. Ces signaux participent non seulement à l'organisation du discours, à l'engagement d'objets et d'agents dans celui-ci mais doivent contribuer aussi à donner des gages (au sens littéral d'engagement) d'attention et d'activité cognitive signifiant à l'autre que son message est compris et intéressant.

Cette communication doit donc être ancrée sur le monde réel : on parle de quelque chose à quelqu'un... dont le système de synthèse doit être conscient : il doit se doter de capacités à construire des représentations de l'environnement dans lequel il est plongé. Pour ceci nous avons besoin d'engager des actions de recherche soit en propre soit en collaboration sur la caractérisation de l'environnement: vision par ordinateur (reconnaissance de locuteurs, identification d'objet...), traitement du signal (séparation parole/non parole, suivi locuteur, spatialisation du son ...). Nous avons pour ceci des collaborations avec plusieurs équipes proches :

- A l'ICP : « Perception » et « Anthropologie Linguistique de la Parole »
- Dans le bassin grenoblois : l'équipe « Prima » du laboratoire GRAVIR, les équipes « Images et Objets » du groupe GOTA et « Multimédia et fonctions associées » du groupe SIC du laboratoire LIS

Dans ce même cadre, j'ai lancé le projet « PRESENCE » qui sera présenté aux fédérations IMAG et ELESA à la rentrée. Ce projet consiste à développer des systèmes de réalité partagée où mondes physique et virtuel sont conscients l'un de l'autre et où les agents de communication respectifs (usagers, clones parlants) peuvent engager des conversations de manière intuitive engageant des objets physiques (objets de l'environnement, par ex. œuvres d'art dans un musée) ou virtuels (par ex. icônes). Une première maquette de type kiosque (intégrant un écran monté sur une tourelle mobile – afin d'assurer un face-à-face – et des microphones et des caméras – afin de combiner localisation acoustique et localisation visuelle ainsi que suivi du regard) est en cours de construction au sein de la salle MICAL, décrite ci-dessous. Elle est à présent couplée avec le système de deixis multimodale (cf. Figure 4) qui fut présenté en démo à sOc'2003.



Figure 5 : .Salle MICAL (à gauche) créée à l'ICP en 2004. Cette salle expérimentale est constituée d'une salle d'étude où sont introduits les sujets et d'une salle de contrôle où opèrent expérimentateurs et complices éventuels. Les deux espaces sont séparés par une vitre semi-transparente. Outre un important matériel audiovisuel nous disposons de deux oculomètres (à droite) et bientôt un système de capture de mouvements temps-réel afin de piloter nos clones par des mouvements prélevés sur un complice humain.

La salle MICAL : les modèles en action

L'ICP est un laboratoire unique au niveau mondial rassemblant autant de compétences autour non pas d'une discipline mais d'un même objet de recherches éclairé par diverses disciplines. Le laboratoire est marqué d'une forte culture expérimentale : les théories sont confrontées aux données par l'intermédiaires de modèles, dont la théorie fixe la structure et que les données paramétrisent. Ces modèles sont donc souvent issus d'expériences in vitro ou in vivo, permettant de collecter un ensemble important de signaux (physiologiques, géométriques, acoustiques, aérodynamiques voire neurophysiologiques) sur un sujet exécutant une tâche donnée explicitement (par ex. par des consignes) ou implicitement (par ex. par la lecture d'un texte ou un rappel de souvenirs émotionnellement forts). Des expériences de perception viennent ensuite confirmer si le modèle issu des données possède bien les propriétés des données elle-même : d'acteur, le sujet devient alors spectateur attentif de ses actions ou de l'action des autres. Or ces deux phases expérimentales sont souvent déconnectées : la boucle perception-action est coupée : or il est certain que l'absence ou la présence de propriétés attendues dans un stimuli synthétique vont conditionner la réponse du sujet et ceci de manière dynamique. Pour explorer cette boucle d'interaction, il faut donc avoir la capacité d'implémenter des modèles temps-réel permettant :

- Soit de perturber la boucle perception-action (par ex. en manipulant des propriétés des stimuli produits par le sujet)
- Soit de simuler la boucle perception-action (par ex. par des expériences en Magicien d'Oz)
- Soit d'interagir directement avec un modèle complet (par ex. le système de deixis multimodale montré en Figure 4)

Nous avons pour ceci monté une salle d'expérimentation dédiée à l'interaction humaine sur le site GARE de l'ICP, dénommée salle MICAL¹. Cette salle complète l'arsenal expérimental de l'ICP déjà doté d'une salle expérimental orientée « mesures » située sur le site CAMPUS. Elle constitue l'une des plate-formes expérimentales permettant d'étudier de nouvelles formes d'interaction humaine, plate-formes mises en réseau au sein du GIS intitulé « PEGASUS »

¹ en l'honneur de l'abbé du même nom, qui en 1779 présenta à l'Académie Impériale des Sciences de Saint Pétersbourg deux têtes parlantes capables de prononcer un certain nombre de phrases. Placées sur un socle à l'intérieur d'un petit théâtre, un dialogue pouvait être engagé entre les deux têtes : "Le Roi donne la paix à l'Europe. La paix couronne le Roi de gloire. Et la paix fait le bonheur des peuples. O Roi adorable père de vos peuples, leur bonheur fait voir à l'Europe la gloire de votre trône.". Un rapport de l'Académie des sciences et signé, entre autre, par Lavoisier et La Place, décrit en ces termes le mécanisme de création de la parole : "Les têtes recouvraient une boîte creuse, dont les différentes parties étaient rattachées par des charnières et dans l'intérieur de laquelle l'auteur avait disposé des glottes artificielles de différentes formes sur des membranes tendues. L'air passant par ces glottes allait frapper les membranes qui rendaient des sons graves moyens ou aigus; et de leur combinaison résultait une espèce d'imitation très imparfaite de la voix humaine."

Références

- Attina, V., D. Beutemps, M.-A. Cathiard and M. Odisio (2004). "A pilot study of temporal organization in Cued Speech production of French syllables: rules for a Cued Speech synthesizer." Speech Communication **44**: 197-214.
- Aubergé, V. (1992). Developing a structured lexicon for synthesis of prosody. Talking Machines: Theories, Models and Designs. G. Bailly and C. Benoît, Elsevier B.V.: 307-321.
- Aubergé, V. and G. Bailly (1995). Generation of intonation: a global approach. Proceedings of the European Conference on Speech Communication and Technology, Madrid: 2065-2068.
- Badin, P., G. Bailly, L. Revéret, M. Baciú, C. Segebarth and C. Savariaux (2002). "Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images." Journal of Phonetics **30**(3): 533-553.
- Bailly, G. and C. Benoît (1992). Talking Machines: Theories, Models and Designs. Amsterdam, North-Holland.
- Bailly, G. and P. Badin (2000). Les 23èmes Journées d'Étude sur la Parole. In Cognito. **19**: 1-2.
- Bailly, G. (2001). Audiovisual speech synthesis. ETRW on Speech Synthesis, Perthshire - Scotland: 1-10.
- Bailly, G., M. Bézar, F. Elisei and M. Odisio (2003). "Audiovisual speech synthesis." International Journal of Speech Technology **6**: 331-346.
- Bailly, G., N. Campbell and B. Mobius (2003). ISCA special session: hot topics in speech synthesis. EuroSpeech, Geneva, Switzerland: 37-40.
- Barbosa, P. (1991). Génération Automatique de la Prosodie du Français. Grenoble - France, Institut National Polytechnique.
- 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.
- Brooks, R. A., C. Breazeal, M. Marjanovic, B. Scassellati and M. Williamson (1999). The Cog Project: Building a Humanoid Robot" in Computation for Metaphors, Analogy, and Agents. Lecture Notes in Artificial Intelligence. C. Nehaniv. New York, Springer: 52-87.
- Brooks, R. A. (2001). "The Relationship Between Matter and Life." Nature **409**: 409-411.
- Cassell, J., J. Sullivan, S. Prevost and E. Churchill (2000). Embodied Conversational Agents. Cambridge, MIT Press.
- Chen, G.-P., G. Bailly, Q.-F. Liu and R.-H. Wang (2004). A superposed prosodic model for Chinese text-to-speech synthesis. International Conference of Chinese Spoken Language Processing, Hong Kong: 177-180.
- Gibert, G., G. Bailly, D. Beutemps, F. Elisei and R. Brun (accepted). "Analysis and synthesis of the 3D movements of the head, face and hand of a speaker using cued speech." Journal of Acoustical Society of America **118**(2).
- Holm, B., G. Bailly and C. Laborde (1999). Performance structures of mathematical formulae. International Congress of Phonetic Sciences, San Francisco, USA: 1297-1300.
- Holm, B. and G. Bailly (2000a). Génération de la prosodie par superposition de contours chevauchants: application à l'énonciation de formules mathématiques. Journées d'Etudes sur la Parole, Aussois - France: 113-116.
- Holm, B. and G. Bailly (2000b). Generating prosody by superposing multi-parametric overlapping contours. Proceedings of the International Conference on Speech and Language Processing, Beijing, China: 203-206.
- Holm, B. and G. Bailly (2002). Learning the hidden structure of intonation: implementing various functions of prosody. Speech Prosody, Aix-en-Provence, France: 399-402.
- Holm, B. (2003). Implémentation d'un modèle morphogénétique de l'intonation. Application à l'énonciation de formules mathématiques. PhD Thesis. Grenoble - France, Institut National Polytechnique: 239 p.
- Keller, E., G. Bailly, A. I. C. Monaghan, J. Terken and M. Huckvale (2002). Improvements in Speech Synthesis. Chichester, England, J. Wiley & Sons, Ltd.
- Morlec, Y. (1997). Génération multiparamétrique de la prosodie du français par apprentissage automatique. Grenoble - France, Institut National Polytechnique de Grenoble.
- Odisio, M. and G. Bailly (2004). "Shape and appearance models of talking faces for model-based tracking." Speech Communication **44**(1-4): 63-82.
- Raidt, S., G. Bailly, B. Holm and H. Mixdorff (2004). Automatic generation of prosody: comparing two superpositional systems. International Conference on Speech Prosody, Nara, Japan: 417-420.
- Vatikiotis-Bateson, E., G. Bailly and P. Perrier (2005). Audiovisual Speech Processing. Cambridge, MA, USA, MIT Press.