Note de lecture

Multimodality in Language and Speech Systems, édité par Björn Granström, David House, et Inger Karlsson, coll. "Text, Speech and Language Technology", vol. 19, Kluwer Academic Publishers, Dordrecht, juin 2002, 256 pages, ISBN 1-4020-0635-7, 89 €.

Les contributions rassemblées dans le présent ouvrage couvrent un ensemble de recherches allant de l'étude de la multimodalité dans la communication humaine, jusqu'à l'implémentation de systèmes multimodaux de dialogue humain-machine. Ces contributions proviennent de conférences données lors de la 7ème école d'été ELSNET et c'est pourquoi le présent ouvrage apparaît davantage comme un échantillon de recherches, que comme un manuel qui viserait la description exhaustive mais moins approfondie du domaine. L'introduction à l'ouvrage, rédigée par les trois éditeurs, présente dans une perspective cohérente l'ensemble des chapitres, qui sont résumés un par un sur une demi-douzaine de pages.

1. La notion de multimodalité

Les tout premiers paragraphes de l'introduction tentent de mettre en place les notions fondamentales du domaine, notamment la 'multimodalité' – et il nous a semblé utile de situer ce terme ici. Une fois reconnue la difficulté de trouver une définition unique, on peut, avec les éditeurs, estimer que « en essence … la multimodalité est l'utilisation de deux ou plusieurs parmi les cinq sens en vue de l'échange d'informations » (p. 1, nous traduisons). Cette définition nous semble toutefois mettre en porte-à-faux la notion de 'multimédia', avec laquelle elle se superpose.

Nous préférons alors considérer, par exemple avec A.D.N. Edwards (p. 74-78), que la communication s'effectue en effet par le biais de nos cinq sens : vue, ouïe, odorat, goût et toucher, ce dernier se divisant en sensibilité tactile, kinesthésie, et équilibre ; seuls la vue, l'ouïe et le toucher sont utilisés en pratique dans les interfaces humain-machine. Le 'médium' désigne alors la nature du canal de communication, lié à l'organe de sens utilisé. La 'modalité' se réfère quant à elle à la forme du contenu communiqué : contenu linguistique (qui peut être parlé, écrit, signé), contenu graphique (plusieurs formes de dessins, graphes, diagrammes), contenu gestuel, etc. Ainsi, un même *médium* peut servir à transmettre différentes *modalités*.

2. Présentation critique des huit chapitres

Le premier chapitre, rédigé par J. Allwood, décrit les différentes possibilités d'utilisation des gestes, postures, et autres indices corporels dans la communication entre individus. Son introduction, très théorique, relie les types d'information que l'on peut communiquer (information indexicale, iconique, ou symbolique) à la façon de communiquer (indication, démonstration, ou signalisation). Puis, l'auteur opère une classification des mouvements corporels : expressions du visage, mouvements de la tête, des lèvres, des bras, posture, etc. – quinze classes au total. Suit alors une classification du contenu qui peut être communiqué par des gestes : états physiologiques, émotions, gestion de l'interaction, et même des informations factuelles (par exemple, 'je ne sais pas'). Le restant du chapitre propose des relations entre ces différentes classifications, avec une analyse théorique du lien entre geste et parole dans

l'interaction humaine. Ce chapitre offre une perspective générale, souvent spéculative (non expérimentale), sur le rôle des gestes dans la communication. Les tentatives de synthèse semblent toutefois insuffisamment exploitées, et on regrette les erreurs typographiques dans quelques tableaux pourtant prometteurs (tableau 2 sans contenu, tableau 3 incomplet).

Le second chapitre (D. McNeill *et al.*) présente une étude expérimentale détaillée du rôle des gestes dans le monologue oral, réalisée à partir d'enregistrements filmés. La définition du protocole d'analyse des gestes, de la parole (aspects prosodiques), et du contenu (aspects discursifs) est minutieusement décrite, permettant ainsi une évaluation des résultats présentés, voire une réutilisation de ces protocoles. L'analyse détaillée d'un enregistrement de 32 secondes portant sur la description d'une maison, notamment par le biais de la transcription annotée, montre que les gestes permettent de découper la narration en unités discursives ('catchments'), qui sont en accord avec le découpage sémantique du discours. De plus, alors que les gestes sont corrélés aussi avec la prosodie, ils permettent un découpage plus fin et plus précis du discours que celle-ci.

Le troisième chapitre (D.W. Massaro) présente une autre série d'expériences de facture psycholinguistique, portant sur l'apport de la vision à la compréhension de la parole, dans ce que l'auteur appelle la 'parole visualisée' ('visible speech'). Les expériences visent à déterminer l'effet perçu par les sujets lorsqu'ils sont exposés simultanément à des stimuli auditifs – par exemple, le son de la syllabe 'ba' ou 'da' synthétisée – et à des stimuli visuels – le visage prononçant 'ba' ou 'da'. L'effet perçu dépend de la combinaison de stimuli. L'auteur propose un modèle probabiliste, fondé sur la logique floue, qui modélise l'intégration perceptive des deux stimuli. Le modèle est testé en comparant ses prédictions aux réponses moyennes des sujets. L'impact d'autres facteurs sur la perception des phonèmes est également discuté, par exemple l'intégration dans un mot ou l'expression du visage qui les prononce.

Ainsi, ce chapitre, tout autant que le précédent, met en évidence l'importance de la multimodalité pour la perception de la parole par l'individu. Les gestes et les mouvements des lèvres apparaissent donc comme des facteurs essentiels. Il n'est pas certain que ces études trouvent une application rapide au domaine de la communication humain-machine, mais inversement, les avancées technologiques – visages ou voix de synthèse, détection des gestes – apportent une aide significative à ces études. Toutefois, la réalisation d'outils d'aide aux personnes handicapées bénéficie plus directement de ce genre d'études (cf. p. 87, les visualisateurs de parole, ou p. 218, le projet Teleface).

Le chapitre rédigé par A.D.N. Edwards (« Interaction multimodale et personnes handicapées ») commence par une bonne synthèse des notions liées à la 'multimodalité'. L'auteur introduit notamment le concept de *conversion* ('mapping') d'une *modalité* de communication vers un *canal* de communication lié à l'un des cinq sens de l'individu récepteur. Par exemple, le langage parlé est par défaut perçu grâce à l'ouïe, mais sa *conversion* par transcription permet une perception visuelle sans trop de pertes d'information (l'intonation est mal rendue, toutefois). Cette même modalité (la parole) peut être aussi *convertie* vers le canal de communication lié au toucher (alphabet Braille). Edwards applique ainsi, de façon très intéressante, le concept de conversion à la réalisation d'outils d'aide pour les personnes ayant un handicap dans un canal de communication. Plusieurs exemples d'outils figurent en fin de chapitre, telle la représentation tactile de partitions musicales ou de diagrammes, développée par l'auteur lui-même.

Le chapitre rédigé par N.O. Bernsen occupe environ un quart du livre (56 pages). Il s'agit d'une synthèse des différentes possibilités offertes à la technique par les modalités de communication, qui sont tout d'abord organisées dans une taxonomie dont les principes de construction sont décrits en détail. Le chapitre se consacre surtout aux représentations mono-modales des 'sorties' ('output') d'un système, qui sont classifiées selon les caractéristiques suivantes : linguistique ou non, analogue ou non, arbitraire ou non, statique ou dynamique. Le 'médium' de transmission peut être visuel, sonore ou haptique (lié au toucher). Les différentes combinaisons de traits, dans ce modèle dit 'génératif', sont explicitées dans un tableau quelque peu laborieux ($2 \times 2 \times 2 \times 2 \times 3$ entrées), puis de nombreux exemples de représentations 'atomiques' sont fournis et commentés. Par exemple, les représentations analogues *et* statiques *et* graphiques *et* sous forme de graphes peuvent être des graphes à courbes, ou à histogrammes, ou en secteurs. Le chapitre se présente ici comme une description des représentations usuelles associées à chaque entrée de la taxonomie. La visée théorique, à savoir la 'Théorie des Modalités', est présente dans l'ambition de définir chaque modalité dans un formalisme unifié. La théorie, implémentée dans l'outil SMALTO d'aide au choix d'une modalité lors de la conception d'un système (*http://disc.nis.sdu.dk/smalto*), est surtout appliquée au langage parlé. L'étude analyse les arguments en faveur du choix de cette modalité qui ont été recueillis dans la littérature. Une brève conclusion portant sur la combinaison des modalités clôt le chapitre.

Cet important chapitre aurait peut-être mérité de figurer en début d'ouvrage, tant ses visées sont classificatoires et descriptives. Son analyse des modalités permet d'unifier un domaine souvent difficile à cerner (notamment à cause de la diversité des médias ou des capteurs utilisés) et de décrire les instanciations présentes ou à venir des représentations. La méthodologie guidant le choix des modalités en fonction du type d'application présente des liens intéressants avec l'évaluation des applications en fonction du contexte d'utilisation. On regrette toutefois la longueur du chapitre, qui fait parfois perdre de vue l'objectif d'ensemble (les exemples et les commentaires pour chaque modalité auraient pu figurer dans une annexe). Plus sérieusement, on peut se demander si au stade actuel la conception d'un logiciel multimodal passe véritablement par une réflexion aussi théorique sur les modalités à utiliser. On peut en douter : souvent le choix des modalités est imposé de façon assez directe par la tâche et les capteurs disponibles. Mais qui plus est, la Théorie des Modalités ne donne pas encore d'indication sur la façon de combiner les modalités, et on peut craindre que l'explosion combinatoire qui apparaît à ce niveau ne limite les ambitions théoriques.

Les trois chapitres restants décrivent des systèmes multimodaux d'interaction humain-machine. Le chapitre de T. Brøndsted *et al.* décrit essentiellement le système CHAMELEON, qui permet à un utilisateur de demander des renseignements sur l'organisation et le personnel d'un laboratoire dont le plan est posé sur une table. Le système est muni d'une caméra pouvant détecter les actes de pointage de l'utilisateur, d'un système de compréhension de la parole, d'un synthétiseur de parole, et d'un pointeur laser. L'architecture interne, fondé sur un tableau noir, fait en réalité intervenir les modules dans un ordre assez prévisible : d'abord le traitement de la requête et des gestes, puis la mise en relation de la requête avec les données disponibles, enfin la synthèse de la réponse et le pointage du laser. En particulier, les dialogues ne semblent pas dépasser la séquence question-réponse. Cette réalisation est significative par l'intégration des différents équipements dans un système fonctionnel, et met en évidence les problèmes liés aux traitements nécessaires pour répondre à des requêtes bimodales par des réponses également bimodales. Par exemple, l'usage des expression déictiques dans la réponse complémente convenablement l'usage du pointeur laser.

Le chapitre de K.R. Thórisson présente un modèle ambitieux intégrant des mécanismes de perception et d'action dans un agent logiciel, matérialisé sur un moniteur et disposant d'une caméra. Cet agent, nommé Gandalf, est une instanciation d'une architecture plus ambitieuse, YTTM (« Ymir turn-taking model » : modèle Ymir pour les tours de parole). Après une synthèse de différentes modélisations des tours de parole dans le dialogue entre humains, l'auteur présente les nombreux modules déclaratifs qui composent le YTTM, et lui permettent de gérer les tours de parole en fonction des mots et des gestes de l'interlocuteur, et d'utiliser ses propres mots et gestes. Ces modules sont organisés en trois couches, et l'une des idées les plus originales du modèle est d'assigner à ces couches des vitesses de réaction variables : la 'couche réactive' est déclenchée environ 2-10 fois par seconde, la 'couche de contrôle du processus de dialogue' environ 2 fois par seconde, et la 'couche du contenu du dialogue' une fois par seconde ou moins. On ne peut résumer ici cette architecture complexe, mais l'agent obtenu permet, selon l'auteur, d'assurer un dialogue multimodal relativement naturel. L'objet du dialogue est la visite d'un système solaire représenté sur un écran. Une expérience pour mesurer l'interactivité du système classe cet agent entre un humain et un chien de compagnie (p. 203). On constate ainsi que les résultats ne sont pas faciles à évaluer, alors même que la complexité de l'architecture décrite suscite quelques interrogations sur son fonctionnement en conditions réelles.

Le dernier chapitre du livre (B. Granström *et al.*) présente un ensemble de systèmes réalisés autour d'un modèle de visage développé au KTH. Le visage de synthèse est capable de reproduire les mouvements du visage humain lors de l'articulation de phonèmes et de mots ('talking face'), ou lors de l'expression de mimiques et d'émotions. Les différents visages construits à partir d'un même modèle paramétrique sous-jacent, qui prend en compte aussi des parties cachées telle la langue, sont utilisés dans plusieurs applications de dialogue humain-machine. Dans le projet Teleface, ces visages aident les personnes malentendantes à comprendre un signal de parole bruité, en lisant les phonèmes sur les lèvres du visage artificiel. L'évaluation de l'apport d'un tel visage à la compréhension montre de façon convaincante son utilité. Dans les systèmes Waxholm et Olga, un personnage animé interagit avec l'utilisateur pour le guider dans différents contextes, et un autre projet étudie l'usage des visages parlants pour l'apprentissage d'une langue étrangère. On constate donc, dans ce chapitre final, une primauté des applications technologiques, aussi bien sur l'étude théorique des humains ou des modalités, que sur l'élaboration d'architectures informatiques complexes.

3. Analyse critique de l'ensemble

Les contributions au présent ouvrage apparaissent en somme comme une série de recherches qui recouvrent plusieurs aspects fondamentaux du vaste domaine de la multimodalité. L'ouvrage ne prétend pas constituer un manuel, mais bien plutôt une introduction par l'exemple – et il s'agit en l'occurrence d'exemples approfondis et solides. Quelques chapitres adoptent une position plus théorisante : le premier, par J. Allwood, le cinquième, par N.O. Bernsen, et dans une moindre mesure le quatrième, par A.D.N. Edwards. Ce dernier, par son équilibre entre l'analyse théorique et l'illustration par l'exemple, est l'un des plus attractifs de l'ouvrage.

On peut regretter, pour ce qui est de la forme générale de l'ouvrage, une certaine impression d'inachevé, due à la présence de coquilles (par exemple dans plusieurs tableaux), et à une harmonisation insuffisante des présentations (la mise en page des tableaux, les bibliographies par chapitre). Une bibliographie générale et un index auraient certainement accru la valeur de

l'ouvrage, ainsi qu'une division plus nette en deux ou trois parties. La qualité graphique des images et des dessins laisse parfois à désirer. Enfin, on peut aussi regretter le retard dans la publication (juin 2002, pour une école d'été de 1999).

Dans la perspective du Traitement Automatique des Langues, on peut se demander dans quelle mesure le contenu du livre est fidèle à son titre, *Multimodality in Language and Speech Systems*. En effet, moins de la moitié de l'ouvrage traite véritablement de systèmes informatiques, alors qu'une large place est dévolue aux expériences psycholinguistiques, dans lesquelles les systèmes informatiques sont parfois un instrument important. Il n'est pas facile d'exploiter les résultats de ces expériences psychologiques pour la conception de systèmes multimodaux de dialogue humain-machine, bien que le chapitre de K.R. Thórisson, par exemple, contienne de nombreuses références aux études du comportement humain relatif aux tours de parole. On constate également une nette séparation entre une vue théorique des modalités, déclinées en de très nombreuses variétés potentielles, et une vue applicative, où seules les plus évidentes sont utilisés : langue (parlée et écrite), représentations visuelles (schémas, graphes, mais aussi gestes et expressions du visage), modalités haptiques (souris). Des périphériques moins courants viennent compléter cette gamme, tels les écrans tactiles ou les pointeurs laser. Ce sont encore une fois les outils d'aide aux personnes handicapées qui manifestent la plus grande variété.

Enfin, toujours en se référant au titre de l'ouvrage, on peut se demander si la multimodalité est simplement la somme de plusieurs modalités, ou quelque chose de plus. Certains chapitres évoquent simplement différentes modalités (par exemple la taxonomie de N.O. Bernsen s'intéresse aux 'sorties' monomodales), mais les chapitres de D. McNeill *et al.* et de D.W. Massaro montrent clairement que chez l'humain, les modalités sont corrélées, et s'enrichissent l'une l'autre. La multimodalité est alors plus que la juxtaposition des modalités, comme le montre, par l'exemple, le chapitre de K.R. Thórisson. La théorie de l'intégration multimodale ne semble pas abordée directement dans l'ouvrage, pas plus que l'évaluation des systèmes d'interaction multimodale, bien que les études présentées fournissent plusieurs éléments en ce sens. Dans l'ensemble, le présent ouvrage nous semble donc proposer d'intéressants exemples de réflexion sur la multimodalité, et permettra aux chercheurs du domaine de s'ouvrir vers des aspects multidisciplinaires, grâce à la diversité des contributions.

Note rédigée par :
Andrei Popescu-Belis
ISSCO/TIM/ETI, Université de Genève
40, bd. du Pont-d'Arve
1211 Genève 4 – Suisse
andrei.popescu-belis@issco.unige.ch
http://www.issco.unige.ch/staff/andrei/