Etude de l'usage de la parole dans les interfaces multimodales

Lian Catinis

Institut National Polytechnique de Grenoble
I.N.P.G

Thèse

pour obtenir le titre de
Docteur de l'Institut National Polytechnique de Grenoble

présentée et soutenue publiquement

par

Lian CATINIS

le 27 Janvier 1998

Etude de l'usage de la parole dans les interfaces multimodales

JURY
M. DOLMAZON Jean-Marc : professeur INPG – GRENOBLE (président)
M. KHIYAMI Sami : professeur à l'Université de DAMAS - SYRIE (rapporteur)
Mme. FAURE Claudi : chargé de recherche CNRS - ENST - PARIS (rapporteur)
M. BISSERET André : directeur de recherche INIRIA – GRENOBLE (examinateur)
M. CAELEN Jean : directeur de recherche CNRS –CLIPS–GRENOBLE (directeur de thèse)

Thèse préparée au sein du laboratoire de Communication Langagière et Interaction Personne – Système (CLIPS-IMAG)

REMERCIEMENT

Je tiens à remercier ici les personnes qui, par leurs conseils et leurs encouragements ont contribué à l'aboutissement de ce travail :

D'abord, Monsieur Jean Caelen mon directeur de thèse qui m'a fait bénéficié beaucoup de son expérience et qui a tout fait pour faire avancer ce travail.

J'adresse aussi un très grand remerciement au Professeur Sami Khiyami qui m'a mis sur une bonne voie au début de ma carrière scientifique et qui continue toujours à m'aider.

Les membres du jury pour m'avoir fait l'honneur d'être rapporteurs et examinateurs et pour leurs remarques et leurs contributions à cette thèse. Je remercie spécialement Monsieur J.M. Dolmazon qui était le directeur du laboratoire ICP où j'ai commencé ce travail, et Monsieur A. Bisseret pour ses conseils chaleureux durant mon travail.

Je remercie mes parents et Georges sans qui, par leurs aide constate je n'aurais jamais pu mener à bien cette thèse.

Et enfin, je tiens à remercier tous mes amis de Grenoble, les membres de l'équipe GEOD au CLIPS et les membres de la société « Systems International » pour leur accueil sympathique et Firas et Mirna pour l'aide qu'ils m'ont apportée pour l'implémentation des logiciels et l'édition de la thèse.

Etude de l'usage de la parole dans les interfaces multimodales

Section 1

Objectifs et introduction théorique

Chapitre 1 : Introduction et objectifs

Chapitre 2 : Définitions, terminologie et état de l'art sur la multimodalité

Annexe 1 : Systèmes multimodaux : état de l'art

Annexe 2 : Bibliographie de la section 1.

Chapitre 1

Introduction et objectifs

La parole est un moyen parmi les plus efficaces qu'utilisent les humains pour communiquer et agir de manière coordonnée. La parole est en général produite avec d'autres signes qui donnent un support au sens à transmettre. Ces signes sont variés : des mouvements des mains et des yeux, des expressions faciales et corporelles, des gestes de désignation, etc.

L'évolution technologique récente des systèmes informatiques rend maintenant possible l'intégration, dans les applications informatiques non seulement du texte et des images fixes mais aussi de l'animation et du son. Il existe également des périphériques spécialisés pour prendre en compte la gestuelle : gant numérique, écran tactile, vision du geste, etc. Donc en même temps qu'augmente la complexité des systèmes informatisés, une croissance comparable des capacités d'échange entre ces systèmes et leurs utilisateurs est en train de s'opérer.

Les récentes avancées dans le domaine de la reconnaissance et de la synthèse de la parole permettent d'envisager la voix comme moyen de communication entre l'homme et la machine. La communication orale soulève cependant des problèmes techniques et ergonomiques particuliers [17], et il ne parait pas souhaitable de proposer des interfaces uniquement vocales, mais plutôt multimodales.

Un système multimodal doit permettre à l'usager d'utiliser le ou les modes d'interaction les mieux adaptés à ses préférences, à son degré d'habileté et à la nature de la tâche à accomplir. Cela implique que le système doit être capable de connaître le moyen de communication le plus efficace pour chaque utilisateur et chaque type de donnée manipulé, et choisir de façon pertinente entre les médias et les modes de présentation de l'information. Il doit aussi être capable de s'adapter aux utilisateurs, du novice au plus expert, en proposant une variété de modes d'expression qui permet à chacun d'utiliser ses propres stratégies de résolution de problèmes et lui donner la possibilité de concentrer son attention sur la tâche à accomplir plutôt que sur la forme de l'interaction.

Pour concevoir de telles interfaces, il est nécessaire de recueillir le maximum de connaissances concernant le traitement de l'information en général par l'homme et plus particulièrement au cours d'un dialogue avec un ordinateur.

Il existe à l'heure actuelle, encore très peu de données expérimentales qui puissent permettre de dégager des règles générales et des concepts stables pour la communication homme-machine multimodale. Nous avons donc choisi d'adopter une approche expérimentale pour mieux cerner les problèmes spécifiques à ce type de communication. Pour cela, nous avons limité notre champ d'étude à la bimodalité parole / gestes de désignation en entrée d'un système graphique.

Les objectifs de ce travail concernent les trois points suivants : l'étude de l'usage de la multimodalité, la réalisation d'un système multimodal et l'étude de l'utilisabilité de la multimodalité.

En ce qui concerne l'usage de la multimodalité nous avons effectué deux expériences sur des sujets de test : d'abord dans un cadre de communication inter-humaine qui simule une interface homme – machine réelle, puis en situation de communication homme-machine réelle avec une interface multimodale. Le but de ces deux expériences est d'observer ce que l'interface multimodale peut offrir et qu'elles sont ses limites.

En ce qui concerne l'utilisabilité de la multimodalité nous avons fait des tests à partir d'une expérimentation d'une interface homme-machine multimodale. Notre étude se base sur un modèle cognitif de l'être humain pour définir et évaluer les interfaces homme-machine.

Le but de notre travail en général est d'avancer vers une spécification d'une interface multimodale qui rende la communication avec la machine plus efficace et moins coûteuse.

Dans le chapitre 2 nous présentons des terminologies et des définitions concernant la multimodalité. Nous présentons aussi quelques architectures de systèmes multimodaux afin de donner seulement quelques idées sur le fonctionnement de ces systèmes.

Dans la section 2 (chapitres 3-4-5-6) nous présentons l'étude expérimentale sur l'usage de la multimodalité.

Dans la section 3 (chapitre 7) nous présentons l'étude expérimentale sur l'utilisabilité de la multimodalité.

Dans la section 4 (chapitres 8-9-10) nous présentons une étude temporelle théorique et expérimentale sur les interfaces multimodales.

Dans la section 5 (chapitres 11) nous faisons le point sur les résultats essentiels et nous faisons quelques propositions pour le prolongement de ces études.

Chaque section est suivie par une liste de références et une bibliographie.

Chapitre 2

Définitions, terminologie et état de l'art sur la multimodalité

1. Définitions et terminologie
2. Les techniques et les dispositifs d'interaction multimodale
3. Taxonomie
4. Problématique des interfaces multimodales

Dans ce chapitre nous faisons une présentation théorique de la multimodalité. Nous présentons aussi la terminologie et les définitions que nous utilisons dans ce travail. Nous présentons en bref les techniques utilisées en interaction multimodale, une taxonomie de la multimodalité, et nous situons la problématique des interfaces multimodales. Dans l'annexe 1 de la section 1 nous présentons quelques exemples de réalisation de systèmes multimodaux. L'annexe 2 de la section 1 contient une liste de références et une bibliographie de la section 2.

1-Définitions et terminologie

La terminologie dans le domaine des interfaces multimodales n'est pas bien stabilisée à l'heure actuelle. Nous présentons ci-après la nôtre afin de fixer quelques définitions (cette terminologie est inspirée de celle de notre équipe et des différents échanges des groupes de travail du GdR-PRC Communication homme-machine).

Média : Synonyme de "médium". Ce terme concerne les dispositifs physiques permettant d'interagir avec le système, en entrée et en sortie et concerne la capture, le stockage ou la présentation des informations [35], [32]. Il est donc défini par le type de capteur ou d'effecteur qu'il met en œuvre [33].

Canal, mode, modalité : ces termes sont équivalents et ils correspondent à la nature des informations, et se réfèrent aux sens humains.

Le terme de modalité a donné lieu à de nombreuses définitions. Selon Bellik[5] une modalité "(…) est définie par la structure des informations échangées telle qu'elle est perçue par l'être humain". On peut la définir également comme un style d'interaction [31] qui peut être utilisé en entrée (orale, gestuelle, écrite…) ou en sortie (sonore, visuelle, tactile…).

Énoncé : un énoncé représente un échange de l'utilisateur à l'adresse de la machine ou de la machine vers l'utilisateur. Il a pour finalité l'atteinte (ou la progression vers) d'un but [49]. Il peut être composé de plusieurs modalités, l'énoncé est alors multimodal.

Les différents composants d'un énoncé doivent être produits dans des voisinages temporels proches et sémantiquement liés. Les auteurs [41], [7], [50] dénoncent le caractère arbitraire de ce critère temporel, qui a déjà montré son insuffisance pour l'association du geste et de la parole dans les premières réalisations multimodales. Zanello[62] pense que ce

critère ne peut être indépendant de l'interprétation, les sujets pouvant faire une pause entre deux parties d'un énoncé, non parce qu'ils désirent donner deux commandes distinctes à la machine, mais parce qu'ils réfléchissent au moyen de formuler l'énoncé. Le critère principal est alors l'unité d'action.

Multimédia : du côté machine, un système multimédia est capable d'acquérir, de stocker, et de restituer des informations de nature différente (texte image, sons, séquences vidéo…) » [5], [36]. On peut différencier les médias dynamiques (audio, vidéo, et animation) et statiques (images, graphiques, textes) [4]. Un système multimédia ne possède pas d'analyseur sémantique de données et traite seulement les données d'un bas niveau d'abstraction, sans processus d'interprétation [37], [17].
Du côté utilisateur, un système multimédia offre à l'utilisateur la possibilité d'acquérir et de manipuler des informations de nature variée (textuelles, graphiques, sonores, etc.), et permet ainsi de rejouer des séquences vidéo, sonores, de consulter des bases de données ou des éléments graphiques [62].

Multimodal : on peut parler de l'interaction multimodale et de la multimodalité ainsi que des systèmes multimodaux.

✗ *Interaction multimodale, multimodalité*

Selon [16], la plupart des interfaces sont multimodales (par exemple, le système Windows95 autorise le "drag and drop" et le raccourci clavier), cependant ce sont les mêmes canaux sensori-moteurs qui sont mis en jeu.
Nous considérons, pour notre part, qu'une IHM est multimodale si elle permet l'utilisation de plusieurs canaux sensori-moteurs (vision, parole, geste, etc.) de manière simultanée, et/ou complémentaire [26].

✗ *Système multimodal*

Le système multimodal ne permet pas seulement de rejouer des séquences de nature diverse (à la différence du système multimédia), il permet aussi la réalisation de tâches de manière interactive avec la machine, pouvant mettre en jeu plus d'une modalité en entrée et/ou en sortie. Il possède un processus de compréhension et traite les différents types de données à des niveaux d'abstractions divers, il échange de manière dynamique ces données avec l'utilisateur.

2- Les techniques et les dispositifs d'interaction multimodale

2-1- Les techniques en interaction multimodale

2-1-1- La reconnaissance vocale

Les critères servant à caractériser un système de reconnaissance vocale sont nombreux. On peut identifier par exemple le mode de communication (direct vs médiatisé, par exemple par le téléphone), le type d'utilisateur (professionnel vs grand public), l'environnement (niveau sonore). Parmi cette variété de critères, trois semblent plus utilisés : le débit de parole que le système est capable de traiter (parole continue/mots isolés), l'étendue du vocabulaire, et le caractère monolocuteur ou multilocuteur du système [1].

La détection de la parole

La parole en entrée peut être détectée de plusieurs manières par le système de reconnaissance :

- il peut être "à l'écoute" et interpréter tout ce qui est dit,
- il peut être déclenché par un mot clef ou par une touche spécifique sur le poste de travail.
- Il peut être déclenché sur le niveau [45], les commandes chuchotées seules étant par exemple, destinées à la machine.

Le débit autorisé

Pour les mots isolés, une pause doit être marquée entre chaque mot. Ces systèmes sont les plus faciles à mettre en œuvre car le problème de l'identification des début et fin de mots est simplement résolu. Ils ont aussi de bonnes performances. Pour l'utilisateur, par contre, ce genre de système demande un entraînement, surtout dans le cas de commandes complexes, car il n'est pas naturel de séparer nettement chaque mot en énonciation non contrôlée (il se pose aussi la question des liaisons entre mots).

Pour la parole continue, les systèmes sont plus délicats à mettre en œuvre et moins robustes. Mais ces systèmes restent plus ergonomiques dans la mesure où aucune contrainte sur l'énonciation n'est requise.

Le vocabulaire

Sa taille influence les temps de réponse, les performances, les besoins en mémoire de la machine, ainsi que le confort de l'utilisateur. Un vocabulaire trop restreint pour la réalisation des tâches peut en effet

entraîner le rejet du système de reconnaissance par les utilisateurs. On distingue généralement quatre grands types de vocabulaires selon leur taille, en ordre de grandeur :

Petit vocabulaire	entre 10 et 100 mots
Moyen vocabulaire	entre 100 et 1000 mots
Grand vocabulaire	entre 1000 et 10000mots
Très grand vocabulaire	entre 10000 et 100000 mots

[62][17]

L'aspect mono vs multilocuteur

- Un système monolocuteur implique un apprentissage de la voix du locuteur et donc un meilleur taux de reconnaissance. Par contre tout nouvel utilisateur doit effectuer cette phase d'apprentissage pour que le système puisse traiter ses entrées.
- Un système multilocuteur permet une utilisation "grand public" et ne nécessite pas d'apprentissage, mais les performances sont moindres.

2-1-2- La synthèse de la parole

Deux techniques différentes sont utilisées, soit le message est préenregistré par morceaux et restitué en insérant des éléments variables, soit on procède par synthèse à partir d'un texte (text-to-speech).

La première technique a l'avantage de présenter une très bonne qualité sonore, puisque provenant d'une voix réelle. Cependant elle doit être utilisée dans les domaines où le vocabulaire et les messages sont parfaitement connus à l'avance, car l'extension du vocabulaire est difficile. On se heurte également à une prosodie non naturelle et à des discontinuités entre les séquences de messages.

La seconde technique présente l'avantage d'un vocabulaire quasi illimité, mais la variabilité prosodique reste faible. Les systèmes de synthèse peuvent reproduire d'une manière très satisfaisante l'intonation humaine, mais elle reste la même d'une fois sur l'autre, entraînant rapidement un sentiment d'agacement. L'introduction d'une variabilité prosodique peut pallier ce défaut [62].

2-1-3- La reconnaissance du geste

Dans l'IHM le geste peut avoir divers aspects et rôles : il peut être décrit comme une extension de la communication orale qui permet d'exprimer

des choses difficilement verbalisables, et qui ne doit pas être envisagé comme un moyen de substitution (au sens de la langue des signes) de la communication verbale [43]. Le geste est souvent réduit à la désignation spatiale en deux dimensions (geste 2D) dans les interfaces graphiques. Il peut dans ce cas être considéré comme très spécifique (puisqu'il ne correspond pas à des situations naturelles) mais il a l'avantage d'être peu ambigu [21]. Le geste 3D, bien que plus naturel, nécessite des moyens de capture sophistiqués. Il est aussi plus ambigu.

Il existe des modes gestuels indirects (qui portent l'information par l'intermédiaire d'un dispositif d'interaction laissant une trace) [39]:

- L'écriture ou le mode linguistique,
- Le symbolique (correspond à tout vocabulaire de signes définis conventionnellement tels que les signes mathématiques, chimiques…),
- Le graphique (regroupe les diagrammes, schémas, tableaux…),
- La désignation (par l'intermédiaire du stylo ou de la souris).

Dans le cadre de la reconnaissance de gestes 3D on peut citer [62] : le gant de données, la caméra qui, en couplant le signal vidéo et un traitement d'image en temps réel, reconnaît les mouvements manuels.

2-2- Les dispositifs d'interaction multimodale

Nous présentons ci-dessous les divers dispositifs qui permettent l'interaction avec la machine, en entrée et en sortie.

2-2-1- Les dispositifs en entrée

On peut classer les dispositifs d'entrée en plusieurs groupes : les dispositifs de pointage (souris, stylo, gant de données...), les claviers, les entrées sonores (microphone), les entrées visuelles (caméra), et les autres (capteurs) [62].

Des études comparatives ont été fait entre les différents dispositifs d'entrée de désignation [14]. Avant de présenter les divers moyens actuels d'interaction, nous rappelons ci-dessous les principaux résultats de ces études.

- La "souris" est la plus efficace en terme de rapidité et de fiabilité pour des sujets expérimentés.
- Le "stylo optique" donne de meilleurs résultats que la "souris" pour des sujets non expérimentés, mais est jugé plus fatigant.
- Le "joystick" est plus efficace que la "souris" pour des déplacements supérieurs à 5 centimètres.

Le "knee-control" est peu précis mais permet de garder les mains libres.

↳ Le clavier

Il permet différents niveaux d'opérations :

- La frappe de caractères.
- L'action (accepter ou rejeter l'action en cours).
- Le changement de mode (majuscule - minuscule par exemple).
- La navigation (avec les touches de déplacement).
- et d'autres actions (les claviers MIDI par exemple pour générer de la musique)

La vitesse de frappe dépend du niveau d'expertise (sujet expert : 1mot/sec ; sujet novice : entre 0.2 et 0.4 mot/sec) et est de deux à quatre fois plus lente que la parole spontanée [53].

L'avantage de ce dispositif est qu'il n'est pas ambigu : le sujet peut ou non avoir effectué la bonne frappe mais il n'y a pas de doute sur ce qu'il a fait [18]. L'inconvénient majeur du clavier est le lourd apprentissage que nécessite son utilisation (cet apprentissage est un obstacle à l'usage de ce dispositif dont pourtant, la place des touches est destinée à l'origine à réduire la vitesse de frappe [2], mais dont la configuration générale ne respecte pas l'angle naturel des poignets). Cet inconvénient est toutefois compensé par le gain de temps ultérieur comparativement à l'écriture (2 à 5 fois plus rapide) [45].

↳ Les dispositifs de pointage

La souris : elle permet un déplacement continu dans le plan et la manipulation directe d'objets. Le mouvement du curseur sur l'écran

correspond à peu près à celui de la souris sur le plan. La souris présente deux modes de commandes : un clic simple (sélection) et un clic double (ouverture d'un fichier, d'une application...). Ce dispositif existe aussi avec retour d'effort.

Les avantages sont la faible place nécessaire sur l'espace de travail pour effectuer les déplacements (repositionnement possible de la souris) et la possibilité de focaliser l'attention sur l'écran sans suivre les déplacements de la souris elle-même. Ce dispositif possède aussi quelques inconvénients. La place nécessaire, même si elle est restreinte, n'est pas compatible avec les portables. Le fait que le bouton de confirmation soit sur la souris, et que celle-ci puisse être en mouvement complique un peu l'activité de sélection sur de petites cibles. La souris fonctionne en mode relatif, ce qui rend son utilisation pour l'écriture et le dessin à main levée très limitée. Ces dernières activités sont d'ailleurs plus naturelles avec un crayon [42].

Le stylo : il peut être utilisé en pointant ou en glissant les objets. Le format et l'espacement des cibles dépendent de la taille du champ de sélection du stylo. Un stylo à large champ permet une sélection plus aisée de petites cibles, mais demande un espace plus grand entre elles pour éviter des erreurs. La précision la plus grande est d'environ 5 pixels.
Les avantages sont l'utilisation d'un même dispositif en sortie et en entrée, et l'utilisation naturelle de l'activité de pointage. Ce dispositif ne prend pas de place sur le plan de travail, et il peut être adapté à des utilisateurs handicapés ou ayant les mains prises, par une fixation du dispositif sur la tête, ceci augmentant cependant la fatigue.

Les inconvénients sont la fatigue qu'engendre ce type de dispositif et l'ajout d'une tâche de sélection (presser et relâcher le bouton de commande du stylo). D'autre part le bras et le stylo peuvent cacher par moment des parties de l'écran. Du fait de la configuration de l'écran, les erreurs sont plus grandes sur les bords de l'écran (parallaxe). Un moyen d'y pallier est de proposer des cibles suffisamment grandes sur les bords pour réduire l'impact de ce problème [42].

L'écran tactile : le nombre de points de contacts possibles varie de 4000*4000 à 25*40 selon la technique utilisée [42]. La facilité d'apprentissage (du fait de l'utilisation naturelle de l'activité de pointage), l'absence de mémorisation des commandes (puisqu'elles sont présentes à l'écran), et l'utilisation du même dispositif en entrée et en sortie sont les avantages essentiels de l'écran tactile. Les inconvénients sont la faible précision pouvant entraîner du désagrément chez le sujet, et aussi le

niveau de pression nécessaire pour l'activation de l'entrée. En effet pour certains écrans tactiles il suffit d'effleurer l'écran, alors que pour d'autres une pression soutenue est nécessaire. Cette variabilité peut apporter de l'inconfort à l'utilisateur. D'autre part, ce dispositif est relativement fatigant (du fait du bras tendu vers l'écran), et indépendamment de la précision des écrans tactiles, la taille des cibles est assujettie à celle du doigt de l'utilisateur. Enfin, on peut noter que l'écran tactile est utilisé principalement pour la sélection d'objets et non pour leur manipulation (à cause principalement du manque de précision de ce dispositif), ce qui limite ses domaines d'application.

La tablette graphique : elle est placée devant ou à côté de l'écran, et le déplacement d'un doigt ou d'un stylo sur sa surface déplace le curseur sur l'écran. Comme le dispositif est indirect, il est préférable qu'il y ait un mécanisme de confirmation (par un indice sonore ou visuel).

L'avantage de l'utilisation du doigt est qu'elle est instantanée (il n'y a pas à se saisir d'un dispositif particulier). Par contre, le stylo permet une meilleure précision (pour la désignation de la cible ou des mouvements fins). Un autre avantage est la reconnaissance de position dans le plan et de pression (par exemple, une pression légère sert à dessiner et une pression plus appuyée permet de détecter une commande). L'utilisation de la tablette graphique est intuitive puisqu'elle utilise les compétences quotidiennes de pointage. La structure même de la tablette en une seule pièce, fait qu'elle peut être utilisée dans des environnements peu compatibles habituellement avec l'outil informatique (environnements poussiéreux en particulier). Avec ce dispositif, il n'y a pas de problème de parallaxe ou de recouvrement partiel de l'écran par le bras ou la main.

L'inconvénient majeur de ce dispositif est qu'un mouvement involontaire peut déclencher une commande. Il ne permet pas une coordination directe entre l'œil et la main, chacune des informations se trouvant dans un plan différent. Pour une utilisation avec le doigt, le frottement continu peut devenir gênant à la longue pour l'utilisateur. La fatigue peut se faire sentir si les sujets doivent garder l'avant-bras et la main suspendus au-dessus de la tablette pour éviter les commandes involontaires. [42]

La boule (trackball) : à l'inverse de la souris, ce n'est pas la main qui bouge, mais la boule, qui déplace le curseur.
Les avantages sont les mêmes que pour la souris. Deux avantages supplémentaires sont le feed-back direct sur les mouvements du curseur, et l'économie de mouvement pour l'utilisateur puisque la main peut rester dans la même position, et que seuls les doigts déplacent le curseur. Le

mode relatif, comme pour la souris, présente les mêmes inconvénients pour le dessin [42]. D'autre part ce dispositif donne de mauvais résultats au pointage et pour le déplacement d'objets comparativement à la souris ou au stylo [48].

Le joystick : il y a une relation continue entre le mouvement du joystick et le déplacement généré à l'écran. Ce dispositif semble plus intéressant pour des tâches graphiques que pour le contrôle de trajectoire. Il peut être aussi utilisé pour les tâches de pointage ne nécessitant pas une grande finesse. Ce dispositif existe aussi avec retour d'effort. Les avantages sont la faible place nécessaire sur l'espace de travail (il peut être intégré à un clavier), et le peu de fatigue qu'il génère [42].

Le gant de données (dataglove) : ce dispositif, contrairement à ceux que nous avons vus précédemment, autorise l'éloignement par rapport à l'écran. Le gant relève positions, orientations, et mouvements de la main en trois dimensions. Des capteurs de flexion identifient aussi la position des doigts. Ce dispositif est surtout utilisé en réalité virtuelle ou réalité augmentée, et permet aux utilisateurs de manipuler des objets virtuels comme ils le feraient avec des objets réels. Ce mode d'interaction correspond aux modes naturels de désignation et de manipulation [42], [47]. Il peut être utilisé en sortie aussi quand le retour d'effort est intégré, et met ainsi en jeu les modalités tactiles et haptiques rarement sollicitées dans l'IHM.

Ces deux derniers dispositifs ont comme justification principale de leur existence, le caractère naturel qu'ils tendent à donner à l'interaction, mais trouvent cependant des critiques dans le fait qu'ils ne permettent pas l'utilisation des deux mains comme il est d'usage dans la vie courante [19].

↳ Les entrées visuelles

Certaines recherches portent sur l'intégration du mouvement des yeux comme dispositif d'interaction avec la machine. Ce moyen d'interaction est intéressant pour la sélection et la détection de cibles, car le mouvement des yeux est indissociable de cette activité. En enlevant des intermédiaires (du traitement de l'entrée visuelle à la commande manuelle) on réduit beaucoup les temps de réponse. Le problème le plus important réside dans les mouvements involontaires des yeux, présents même si le sujet est très concentré. Cette technique implique aussi le port d'un équipement spécial (caméra qui analyse les mouvements des yeux),

nécessite un apprentissage, et son coût est élevé [42]. Il faut de plus que les objets soient présents à l'écran.

À cause de la place en mémoire prise par ce type de données, on se sert des images directement sur l'écran de la machine ou on effectue une compression pour les rejouer plus tard. La caméra peut servir d'intermédiaire à la lecture des lèvres ou à la visiophonie.

↳ Les entrées orales

Elles se font au moyen du microphone, d'une carte vocale et d'un disque magnétique qui permet le stockage de données et leur restitution ultérieure. Ce stockage sert au système de reconnaissance, et peut permettre un filtrage sur le niveau et la qualité sonore.

L'utilisation de la parole dans l'IHM présente des avantages pour l'utilisateur car elle est un canal naturel de communication, et son utilisation ne nécessite pas, a priori, d'apprentissage particulier. Elle peut être utilisée de façon indépendante ou en complément à d'autres modalités. Elle est particulièrement utile pour des personnes handicapées, ou dans les situations où les mains sont déjà requises pour une autre tâche.

Les problèmes s'opposant à son utilisation en IHM sont relativement nombreuses, et concernent à la fois le niveau technique et humain. Il y a les limites que connaissent encore actuellement les systèmes de reconnaissance tant au niveau du vocabulaire, des grammaires (problème du traitement de la parole spontanée, des ambiguïtés...) que de la robustesse des systèmes (taux de reconnaissance) ou encore de la variabilité inter et intra-individuelle, ou de l'environnement sonore, etc.
Une autre limite est le délai de réponse des systèmes de reconnaissance. C'est un facteur très important pour l'acceptation de tels systèmes. Des études [28] montrent en effet que la tolérance des sujets dépend de leur niveau d'expertise et de facteurs situationnels tels que le contexte et les contraintes liées à la tâche, et aussi la rapidité de réaction…
Indépendamment de l'aspect technique, l'aspect social est lui aussi important. Pour des applications "grand public" par exemple, l'effet facilitateur de la parole n'est pas garanti car il peut y avoir une gêne pour l'entourage (nuisance sonore) et l'utilisateur (crainte du jugement social en cas d'hésitation, d'erreur…) [18]. Frese [40] fait la différence entre les erreurs en modalité orale et les erreurs en modalité manuelle, ces dernières étant plus discrètes, et moins caution au jugement social. Pour

d'autres applications telles que la bureautique, il peut y avoir perte de confidentialité (le sujet pouvant être entendu par un tiers)

2-2-2- Les dispositifs en sortie

↳ Les sorties graphiques

Elles sont portées par l'écran, dispositif de sortie le plus courant, puisque le premier en date. Il permet de présenter les informations graphiques, statiques ou dynamiques, et textuelles. Ce dispositif, par la permanence de l'information, permet une bonne discrimination par balayages, retours en arrière… La lecture est aussi beaucoup plus rapide que l'écoute (700 mots/min. pour la lecture et 180 mots/min. pour l'écoute)[62]. Les informations présentes sur un écran peuvent être complexes. Certaines résultats en psychologie cognitive [12], [13], [30], [51] permettent d'établir des règles concernant la présentation de textes, de figures, ou de vidéo de manière à ce que la compréhension et la mémorisation des utilisateurs soient optimisées.

↳ Les sorties sonores

Elles peuvent être composées de sons possédant une valeur sémantique pour l'utilisateur ou de messages oraux [62].

Les sons : Ils peuvent être divisés en trois groupes [34] :

- Les alarmes et les avertissements (sont prioritaires sur les autres informations, et ont un caractère interrompant),
- Les messages d'état et de contrôle (permettent d'étendre et d'enrichir l'IHM. Ils peuvent être utilisés pour permettre l'accès à l'ordinateur à des mal voyants, comme par exemple l'interface "SoundTrack" pour le traitement de texte utilisant des sons musicaux et parlés),
- Les messages codés (servent principalement à présenter des données numériques sous forme de sons complexes, en addition ou substitution à des graphiques).

L'utilisation des sons significatifs permet d'annoncer des événements spécifiques (tels que l'arrivée de courrier électronique, l'éminence d'un message vocal...). Ils sont même à préférer à des messages vocaux dans le cas de messages courts [5]. Les temps de réaction sont plus courts pour des messages d'alarmes sonores que pour des messages d'alarmes visuels[62]. Ils peuvent cependant demander un apprentissage important

pour être mémorisés, et en situation de stress, le sujet a tendance à oublier leur signification [57].

La synthèse de la parole : la sortie vocale permet à l'utilisateur une liberté de mouvement qu'il ne peut avoir dans le cas d'une sortie visuelle. Si le résultat d'une recherche est donné de manière sonore, le sujet n'est pas obligé de rester devant son écran. Elle permet aussi de ne pas surcharger l'écran d'informations contextuelles [45], et est très utile pour les non-voyants ou pour pallier de mauvais éclairages.

L'interprétation du message sonore doit être simple, doit pouvoir être répétée à la demande de l'utilisateur, et doit requérir une action immédiate comme c'est le cas des messages d'alarme. Si ce n'est pas le cas, le message sonore devient un obstacle à la réalisation de la tâche en cours et non une aide (car il demande la mémorisation du message ainsi que son moment d'occurrence, et peut aussi interrompre la tâche de manière intempestive) [33].

Les inconvénients sont relativement nombreux, le principal étant la nuisance sonore pour le voisinage. Comme nous l'avons dit plus haut, la production orale (et sa compréhension) est plus lente que l'affichage (et sa lecture). D'autre part une sortie vocale requiert un effort d'attention soutenu de la part de l'utilisateur. Du fait de ce niveau d'attention nécessaire, les messages courts peuvent ne pas être pris en compte, et les messages longs peuvent ne pas être mémorisés dans leur totalité. Ils doivent alors être précédés d'une séquence sonore indiquant l'éminence d'un message oral. L'introduction systématique d'une séquence sonore pour les messages courts, ajoutée à la monotonie des messages synthétisés risque cependant d'accélérer l'agacement des sujets vis à vis du système (Cf. p21).

↳ Les dispositifs de retour d'effort

Dans le cas de manipulation d'objets, la coordination du geste avec la vision peut être nécessaire, par l'intermédiaire d'un système de retour d'effort [5]. Il y a interdépendance entre les informations en entrée et en sortie dans la coordination motrice, ce qui rend très difficile la manipulation d'un objet sans avoir de retour d'information sur ses caractéristiques physiques (pour un gant de données par exemple) [48].

3- Taxonomie

L'interaction multimodale permet l'utilisation de plusieurs canaux sensori-moteurs de manière parallèle ou alternée afin de donner la possibilité de réaliser une même action de plusieurs manières ou plusieurs actions en même temps dans des modalités différentes [62]. L'établissement d'une taxinomie est donc important, pour pouvoir classer ces possibilités selon différents critères (temps, type d'information porté par chaque modalité…), et dans des optiques diverses (conception, évaluation…). Cette classification est double selon que l'on se place du côté de la conception [17], [6] (de la machine vers l'utilisateur) ou de l'utilisation de l'interface (de l'utilisateur vers la machine).

✗ *De la machine vers l'utilisateur*

Bersen [9], [11], [8] a identifié 5 propriétés de base pour caractériser les modalités selon leur pouvoir d'expression :
Représentation analogue / non analogue (la représentation analogue se distingue de manière quasi intuitive de la représentation non-analogue dans la plupart des cas. La représentation analogue est appelée aussi isomorphique ou iconique. Elle est spécifique, mais n'a pas de focus, et ne permet pas non plus une grande variété d'interprétation de ce qui est représenté [62].)
Représentation arbitraire / non arbitraire (présente une charge supplémentaire pour l'utilisateur car elle nécessite un apprentissage des conventions de représentation. Cette représentation est liée la plupart du temps au caractère analogue vs non analogue. Les graphiques standard par exemple sont non analogues et arbitraires.),
Représentation statique / dynamique (distinction importante car l'une ou l'autre représentation ne va pas avoir la même pertinence selon le contexte de la tâche),
Représentation linguistique / non linguistique (la représentation linguistique manque de spécificité, contrairement aux représentations analogues, mais permet de faire le focus sur l'information.),
Représentation selon le média, et la structure de la modalité (on peut distinguer le média d'expression correspondant au toucher, à la vision et à l'audition [8]. La même information linguistique peut donc être portée par l'un des trois médias, mais la pertinence de l'un ou l'autre va dépendre du but à atteindre).

D'autres recherches [23], [56] établissent une classification des systèmes selon le type de multimodalité qu'ils permettraient de traiter, et

distinguent quatre grands types intégrant la notion de temporalité (les propriétés CASE).
La multimodalité concurrente : plusieurs énoncés correspondant à plusieurs actions sont produits en même temps en utilisant des modalités différentes.
La multimodalité alternée : une seule modalité est utilisée à un instant donné pour une action, mais plusieurs sont utilisées pour une même tâche.
La multimodalité synergique : plusieurs modalités sont utilisées par énoncé et en même temps.
La multimodalité exclusive : une seule modalité est utilisée à un moment donné, et pour une tâche.

Nous allons nous fonder sur ces propriétés générales pour classer les systèmes multimodaux et tout d'abord nous présentons les propositions de différents chercheurs.

Bellik [6] a repris dans sa classification ces grandes classes, et a retenu trois paramètres :
1. la production des énoncés (séquentielle, parallèle)
2. le nombre de médias dans un énoncé (un, plusieurs
3. l'usage des médias (exclusif, simultané)

A partir de ces trois paramètres il identifie sept types de multimodalité. On peut remarquer que les deux premiers paramètres sont liés à la production de l'utilisateur, alors que le troisième dépend de la capacité du système à traiter les données. Cette classification se base sur une analyse quantitative et temporelle des données, elle ne prend pas en compte l'aspect sémantique des énoncés [62].

Multimodalité exclusive : une seule modalité est disponible à un instant t. Les requêtes doivent être faites de manière séquentielle.
Multimodalité alternée : une commande peut être réalisée avec plusieurs modalités mais celles-ci doivent être utilisées alternativement.
Multimodalité synergique : une commande peut être réalisée avec plusieurs modalités utilisées simultanément.
Multimodalité parallèle exclusive : plusieurs commandes peuvent être réalisées en parallèle, mais chaque commande doit être réalisée avec une seule modalité, et une seule modalité peut être activée à un instant t.
Multimodalité parallèle alternée : plusieurs commandes peuvent être données en même temps, plusieurs modalités peuvent être utilisées pour une commande, mais une seule modalité peut être activée à un instant t.
Multimodalité parallèle synergique : plusieurs commandes peuvent être données en même temps, elles peuvent comporter plusieurs modalités, et

les modalités des différentes requêtes peuvent être utilisées en même temps.
Multimodalité parallèle simultanée : plusieurs commandes peuvent être données en même temps, elles doivent être composées d'une seule modalité, et les modalités des différentes requêtes peuvent être utilisées en même temps.

Ces classifications "machine - utilisateur" cherchent à recenser les différentes possibilités de production multimodales qu'un système pourrait avoir à traiter, mais le font souvent sans prise en compte des capacités d'utilisateurs potentiels [62].

✗ *De l'utilisateur vers la machine*

L'usage de la multimodalité vu de l'utilisateur est souvent caractérisé sous le terme de "propriétés CARE" [35]. Cette classification fait le lien entre les notions d'état (ensemble de propriétés caractérisant une situation), de but (état à atteindre), de modalité, et de relation temporelle (les diverses parties d'un énoncé se trouvent ou non dans la même fenêtre temporelle). Cette classification est valable pour une caractérisation des activités de l'utilisateur vers la machine. Quatre types d'utilisation de la multimodalité sont définis :

*La **C**omplémentarité :* lorsque pour une action, l'usager utilise une expression multimodale dans laquelle chaque mode est nécessaire (et contribue) à la compréhension de l'action,

*L'**A**ssignation :* lorsque pour une action, l'usager choisit un mode récurrent particulier (ou un sous-ensemble de modes) pour s'exprimer (il se peut aussi que ce choix soit imposé par le concepteur du système), (appelée aussi spécialisation),

*La **R**edondance :* lorsque pour une action, l'usager utilise simultanément deux ou plusieurs modes à travers lesquels les informations sont redondantes,

*L'**E**quivalence :* lorsque pour une action, l'usager choisit indifféremment tel ou tel mode (ou un sous-ensemble de modes) pour s'exprimer. les différentes modalités permettent d'obtenir le même résultat (équivalence de résultat), et peuvent être d'un coût cognitif équivalent (équivalence fonctionnelle) [44].

Une extension des propriétés CARE a été proposée par Zanello, Caelen et Bisseret [61]: les propriétés T-CCARE. Le but de cette extension est de caractériser la réalisation de la tâche.
Dans [61], une classification des énoncés et des tâches en environnement multimodal a été proposée afin de rendre compte des différents usages de la multimodalité et permettre une évaluation et un retour sur la conception. A cette fin, les propriétés CARE ont été complétées d'une part, et d'autre part, elles ont été intégrées avec les propriétés CASE.

Zanello et al. [61] ont remarqué que les quatre éléments cités précédemment (complémentarité, assignation, redondance, équivalence) ne sont pas de même niveau. La complémentarité et la redondance peuvent caractériser l'énoncé multimodal lui-même, sa structure, les éléments informationnels et les modalités qui le composent. Ces propriétés contribuent à l'analyse des énoncés multimodaux, leur aboutissement pouvant être la formulation de recommandations ergonomiques.
L'assignation, et l'équivalence se placent dans une autre dimension. En effet elles s'observent dans le temps pour un type d'objet, ou d'action. Elles résultent de la compilation des observations menées lors de l'utilisation d'une interface et ne peuvent caractériser un seul énoncé.

Dans une expérimentation, Zanello a remarqué que les informations contenues dans chaque modalité peuvent être en conflit et il est nécessaire que le système l'identifie comme tel pour qu'il puisse le gérer et prendre les décisions adéquates.

L'énoncé multimodal peut être de trois types différents (complémentaire, redondant ou conflictuel), et chaque énoncé pouvant être actionnel ou informationnel :

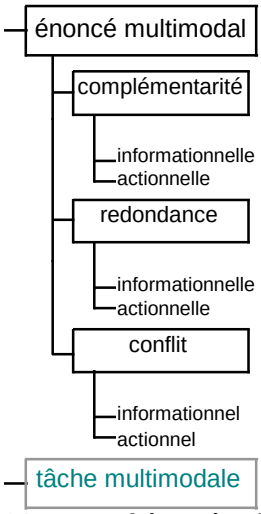


fig1 Extension des propriétés CARE pour l'énoncé multimodal : les propriétés CCARE

L'énoncé multimodal informationnel est un énoncé multimodal lors de sa production tandis que l'énoncé multimodal actionnel est multimodal lors de son exécution.
L'étude d'un énoncé multimodal permet l'étude fine des phénomènes intervenant lors de sa production, mais l'utilisation des différentes modalités au cours d'une tâche peut varier à chaque énoncé. Pour réaliser son but (sa tâche) l'utilisateur peut enchaîner les commandes les unes après les autres ou bien les entrelacer, changer de modalité suite à un échec, ou encore changer de stratégie selon le type de commande. Pour rendre compte de ces différents phénomènes et permettre une analyse quantitative, il est nécessaire d'étudier l'usage de la multimodalité sur un axe temporel.
Ces remarques amènent à inclure les propriétés CCARE portant sur des énoncés isolés, dans un schéma plus large concernant la tâche et réhabilitant la temporalité. Cela conduit au schéma suivant qui définit les propriétés T-CCARE.

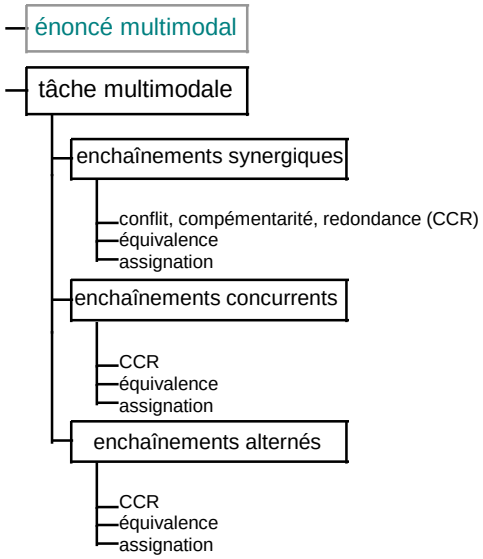


fig 2 Extension des propriétés CARE pour la tâche multimodale : les propriétés T-CCARE

Cette classification doit permettre de quantifier finement les actes multimodaux, de les caractériser, ainsi que d'analyser les différentes stratégies de gestion de la multimodalité lors de la réalisation de la tâche, dans différentes situations expérimentales.

4- Problématique des interfaces multimodales

Les problèmes qui distinguent les interfaces multimodales des interfaces classiques naissent de la diversité des modes en entrée et en sortie dont il

faut analyser, interpréter et générer les informations de manière croisée et dépendante. Ces problèmes concernent [25]:

- La gestion des modes aux niveau des événements (chronologie et synchronie), des informations (unités et actes) et du contexte interactionnel.
- La fusion/fission des informations au niveau morphosyntaxique, sémantiques et/ou pragmatique et actionnel (intégration de la multimodalité au niveau de la couche interaction/dialogue).
- L'échange des informations avec les autres modules de l'interface et le noyau fonctionnel de l'application.

En se limitant aux problèmes des entrées on peut voir les couches de traitement dans les interfaces allant d'un niveau concret, les signaux, à un niveau abstrait, le déclencheur de l'action. Ces couches sont :

- L'acquisition des informations fournis par l'usager,
- leur reconnaissance automatique,
- la compréhension des signes qu'ils véhiculent,
- leur interprétation coréférentielle,
- la construction d'un message actionnel multimodal.

Le cheminement des informations passe par une mise en forme, une représentation abstraite, une fusion et enfin une transmission à la couche « dialogue » qui se trouve de fait posé au niveau plus haut.

Dans l'annexe 1 nous allons présenter brièvement quelques exemples d'architecture et de systèmes multimodaux existants [22].

Section 1
Annexe 1

Systèmes multimodaux : état de l'art

1- Architectures logicielles des systèmes multimodaux d'interaction homme-machine

Cette annexe présente des architectures de quelques systèmes d'interaction homme-machine multimodaux.

Architectures logicielles : Trois types de modèles d'architecture sont utilisés dans la conception des systèmes interactifs.

1-1- Le modèle « Seeheim » : Seeheim est un modèle langagier composé de trois modules (figure 1)

- l'interface avec l'application qui est responsable de la traduction des expressions en expression compréhensibles dans le domaine de l'application.
- le contrôleur du dialogue qui correspond au niveau syntaxique de l'interface et qui est responsable de la vérification de la cohérence, de la bonne construction et de la correction des expressions.

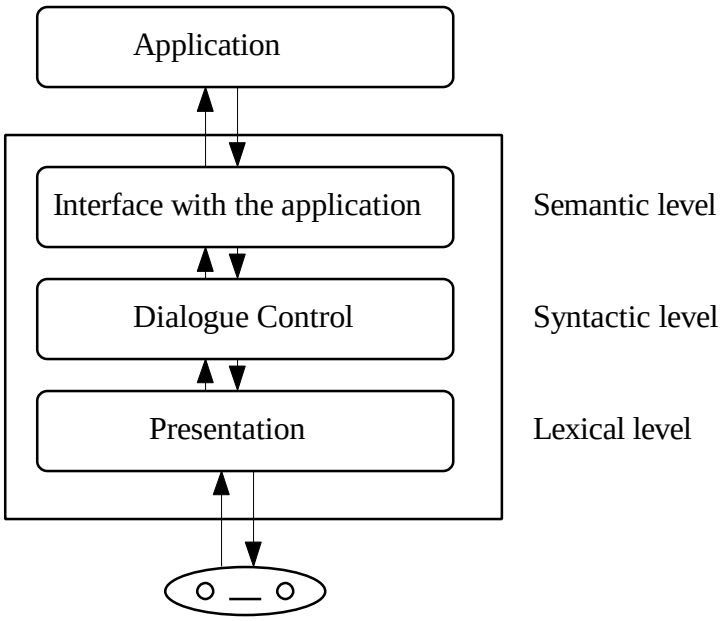


Figure 1 – le modèle Seeheim

- Le module de présentation qui correspond au niveau lexical. Son rôle et de traduire l'expression reçue du contrôleur de dialogue pour la présenter sous la forme appropriée aux médias de sortie. Ce module transmets aussi les données reçues du média de l'entrée au contrôleur du dialogue après les avoir traduites en unités lexicales.

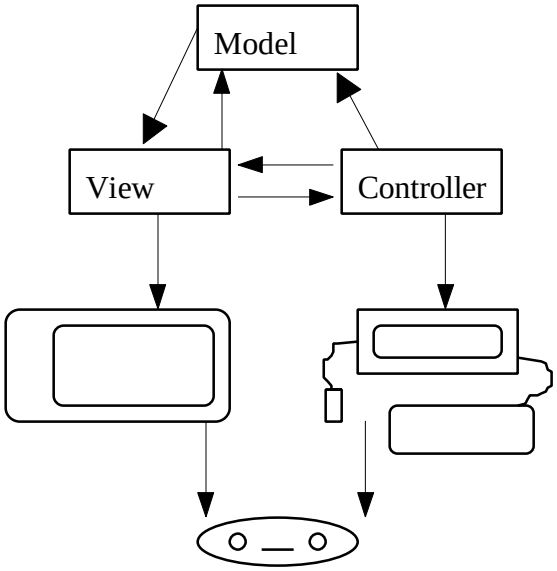


Figure 2 – le modèle MVC (Model View Controller)

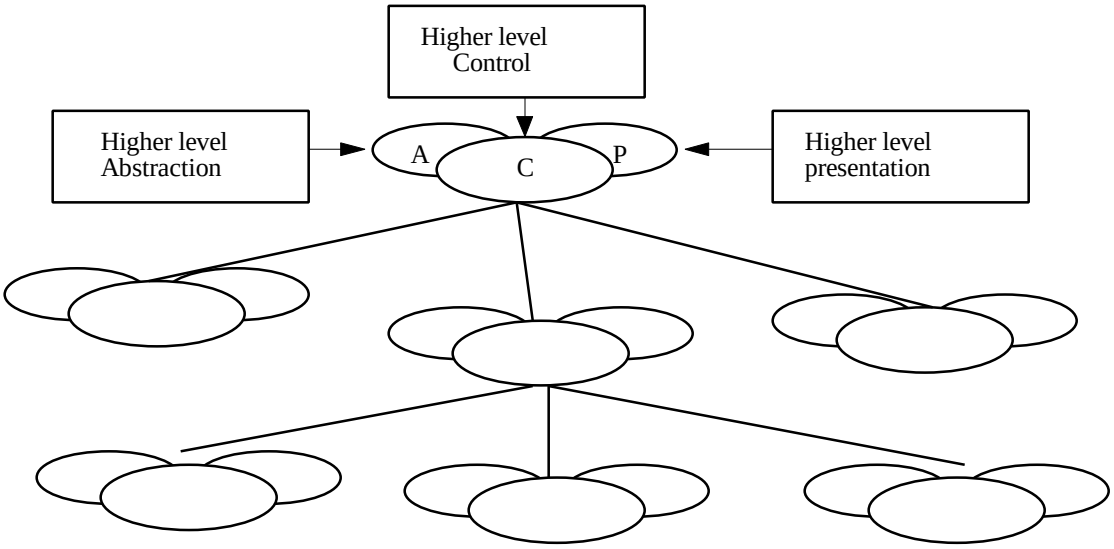


Figure 3 – le modèle PAC (Presentation Abstraction Control)

1-2- Le modèle Multi-agent: le modèle multi-agent est composé d'un ensemble d'agents spécialisés qui réagissent et génèrent des

informations qui correspondent aux événements. Les agents sont des unités autonomes coopératives qui constituent l'interface homme-machine. L'agent a des récepteurs et des transmetteurs avec une mémoire à double niveau (pour mémoriser les événements et les états), et une unité spécialisées de traitement. Plusieurs types de modèles multi-agents sont dérivés du modèle MVC (Model View Controller) (figure 2), et du modèle PAC (Presentation Abstraction Control) (figure 3), qui est structuré en forme hiérarchique et qui est capable d'exprimer les relations inter-agent et la continuité des niveaux d'abstraction.

1-3- Le Modèle à couches : dans ce modèle, l'interface est constituée de plusieurs couches d'abstraction (figure 4). Le programme client a accès à des fonctions de n'importe quelle couche.

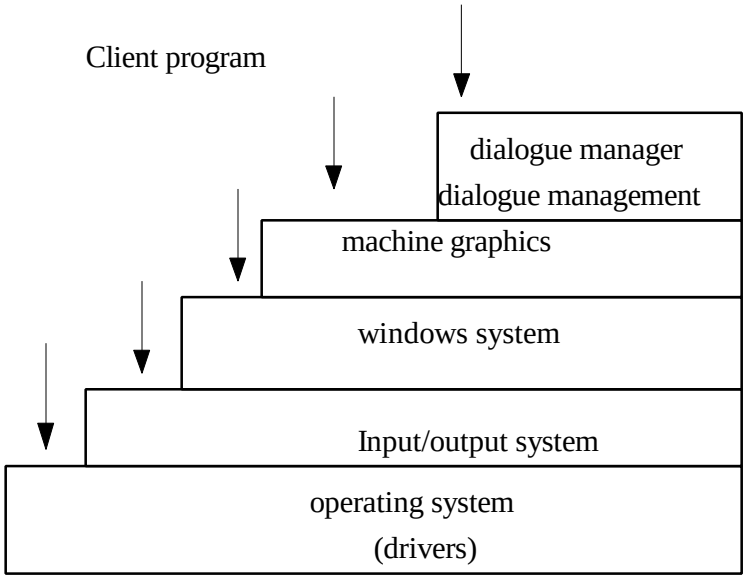


Figure 4 - The layers model

En général, le problème principal est de séparer le noyau de l'application de l'interface utilisateur pour adopter une conception modulaire.

Le modèle à couches peut être utilisé en conjonction avec le modèle multi-agent pour dépasser les contraintes du système d'exploitation, et les contraintes du matériel et des pilotes des médias. La figure 8 présente une architecture générale d'un système de dialogue multimodal [17].

2- Implémentations des systèmes multimodaux

2-1- ICPDRAW

ICPDRAW [27] est une application de dessin multimodale, guidée par des objets, son contexte interactionnel est synergique. Développé à l'ICP-Grenoble (Institut de la Communication Parlée) et repris au laboratoire CLIPS, ICPDRAW est organisé autour d'une architecture distribuée attachée au système UNIX (figure 6). L'échange de données est synchronisé entre le processus client et le processus serveur. L'architecture logicielle inclut le noyau fonctionnel, les niveau de fusion, les gestionnaires des modes, et le serveur multimédia.

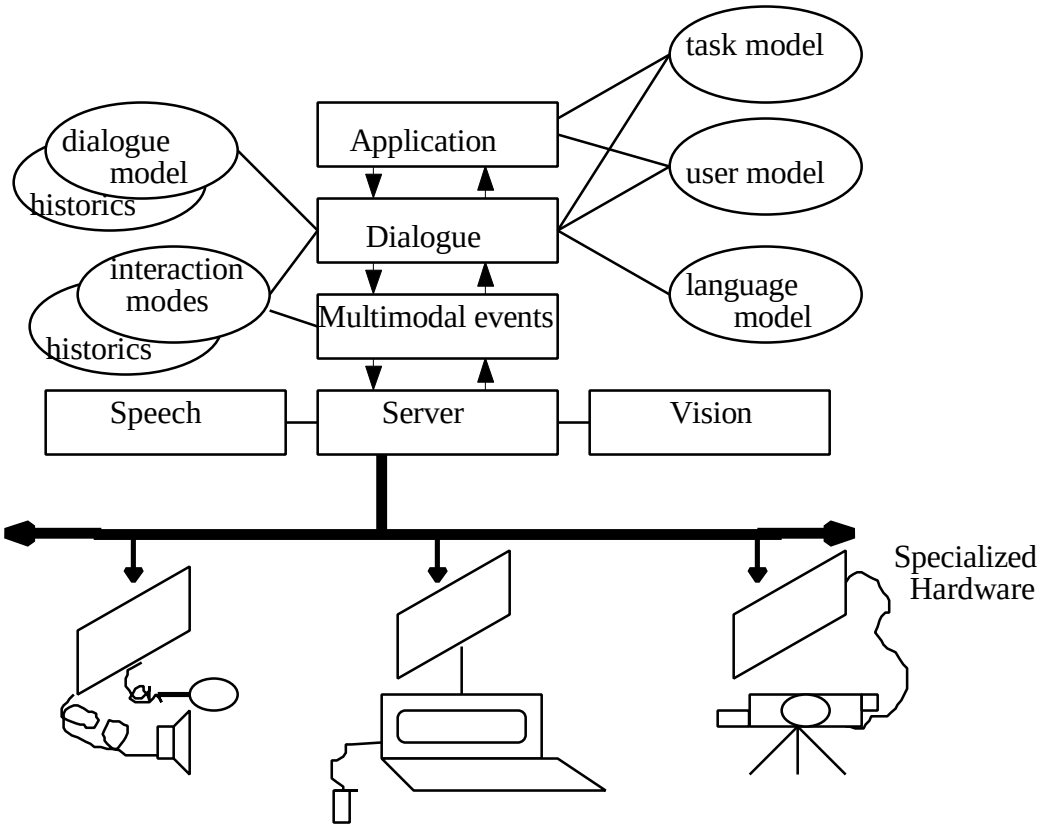


Figure 5 – L'architecture générale du système de dialogue ICPDRAW

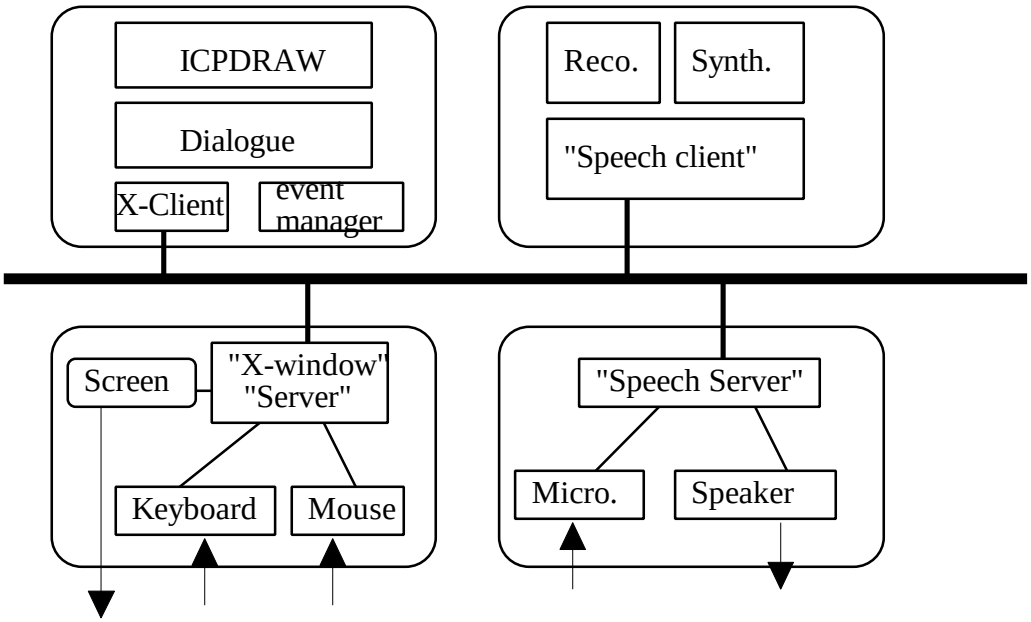


Figure 6 – L'architecture distribuée de ICPDRAW

Les modules du système sont des processus UNIX qui communiquent sous le protocole (ipc). La gestion du clavier et de la souris est faite par un standard X-Windows, la gestion de la parole est faite par le serveur de la parole « Speech Server » qui tourne en arrière plan. Les différe nts modules peuvent tourner sur plusieurs stations et communiquer par le réseau. Les événements sont manipulés par « un tableau noir » qui mémorise les événements, les unités, les actes et les expressions CMR (Common Meaning Representation). La figure 7 présente le diagramme fonctionnel général de ICPDRAW.

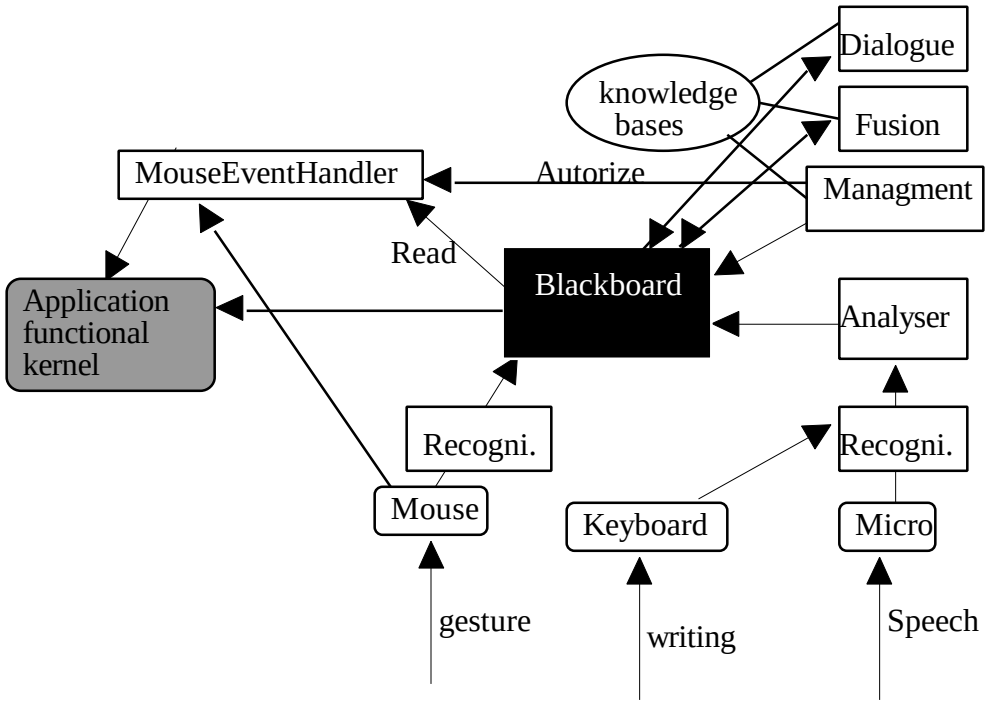


Figure 7 – Echange des informations dans ICPDRAW

2-2- Le système MMI2 (Man Machine Interface for Multimodal Interaction with Knowledge Based System)

L'objectif de MMI2 [59] est de développer un « toolkit » et une méthode pour utiliser ce toolkit dans le but de construire un système d'interrogation d'une base de connaissance dans différents domaines. L'architecture de MMI2 peut être décrite par les trois couches du modèle de Seeheim : la présentation (gestion d'entrée/Sortie), le contrôleur du dialogue et la couche de l'application avec sa base de connaissance. L'architecture est modulaire et elle est composée de "communautés expertes" et chaque niveau est composé d'un ensemble de modules spécialises qui exécutent une tâche spécifique. Chaque module a sa propre structure de données ce qui permet de grouper plusieurs processus cohérents dans un module. La figure 8 illustre l'architecture du système MMI2.

Toutes les opérations dans la couche de gestion de dialogue sont faites en utilisant des CMRs (Common Meaning Representation) qui sont indépendants de l'application ainsi que du mode spécifique. La fonctionnalité principale de chaque module expert est la suivante [59]:

- Le contrôleur de dialogue « *dialogue Controller* » assure la gestion de la fonction de dialogue, formule la CMR et manipule les anaphores et les déictiques.
- Le module «*dialogue context expert*» assure la gestion d'enregistrement de la structure de dialogue et assure l'extraction des informations utiles.

- Le module «*user modeling expert*» assure l'extraction des préférences de l'utilisateur de la CMR et enregistre le modèle de l'utilisateur et offre l'accès à ce modèle.
- Le module «*informal domain expert*» est responsable du stockage des plans de tâche du domaine et évalue les aspects informels du dialogue.
- Le module «*communication planning expert*» est responsable de la construction de la forme logique de la sortie du système et choisit le mode de la sortie.
- Le module «*formal domain expert*» est responsable de la traduction de la CMR en termes du domaine qui représentent les meta-connaissances du domaine.
- Le module «*interface expert*» est responsable de la traduction de la CMR pour l'implémentation physique courante de l'interface.
- Le module «*semantic expert*» a des connaissances autour des définitions intentionnelles des prédicats utilisés dans la CMR.
- Dans le «*mode layer*», les différents modules sont responsables de l'entrée et de la sortie de chaque mode vers la CMR.

Il est clair que l'architecture est modulaire et qu'elle permet d'exporter des informations de n'importe quel module vers un autre système.

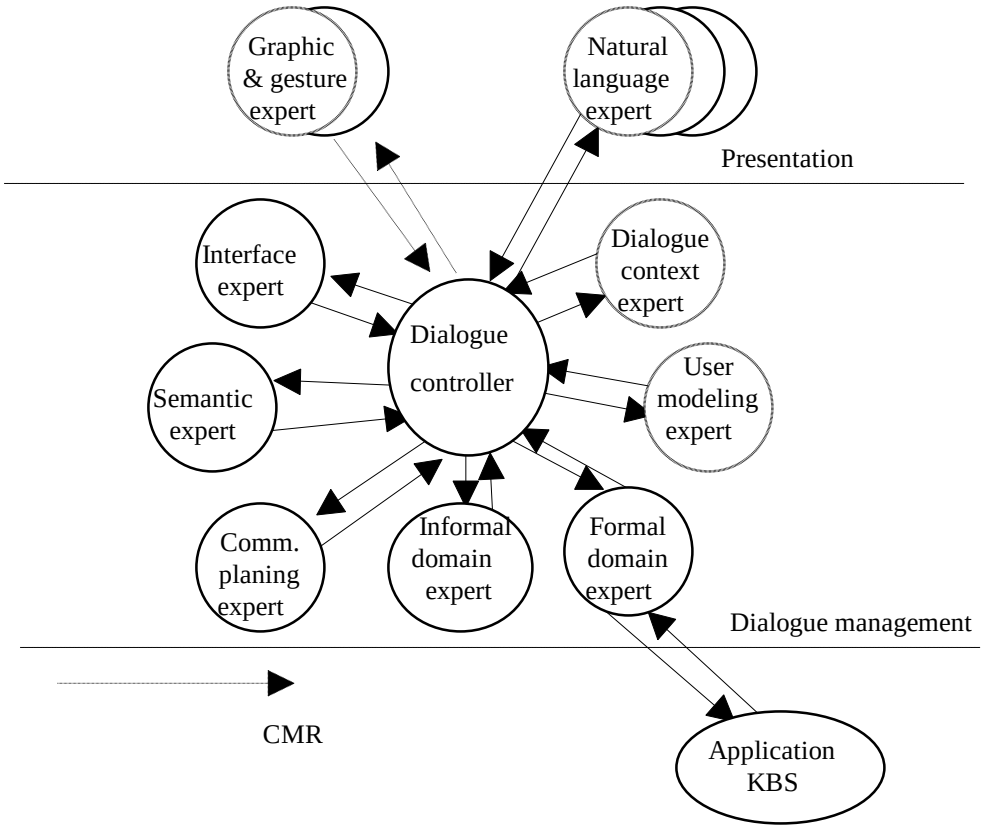


Figure 8 - MMI2 architecture

2-3- MELODIA (**M**ultimodal **E**nvironment for a natura**L** and task **O**riented **DIA**logue)

MELODIA [15] est développé par «Thomson» en collaboration avec le CRIN "Centre de Recherche en Informatique de Nancy". MELODIA est un système général intermédiaire de dialogue multimodal qui prend place entre l'application et l'interface utilisateur.
Les modules principaux du système sont (figure 9):
- Le *media manager*: qui traduit les événements reçus et les différents médias en représentation abstraite indépendante du media et acceptable par l'analyseur. Le gestionnaire de media pilote le media de la sortie et lui transmet les événements construits par le synthétiseur.
- Le *multimodal analyzer* construit une représentation conceptuelle de l'acte multimodal généré par les différents événements, en utilisant une approche qui s'appelle "by lexicon" pour assurer la généralité de l'architecture.
- Le module *dialogue coordinator* effectue la tâche de résoudre d'ambiguïté, des anaphores et des ellipses. Il modifie les différentes bases de connaissance (la hiérarchie des concepts, l'historique du dialogue, le modèle d'activité et le modèle du canal de communication). Ce module fait le lien avec l'application et l'opérateur : il transmet la requête de l'opérateur, la traduit en description fonctionnelle pour le module interface avec l'application et il transmet la requête de l'application à l'interface utilisateur.

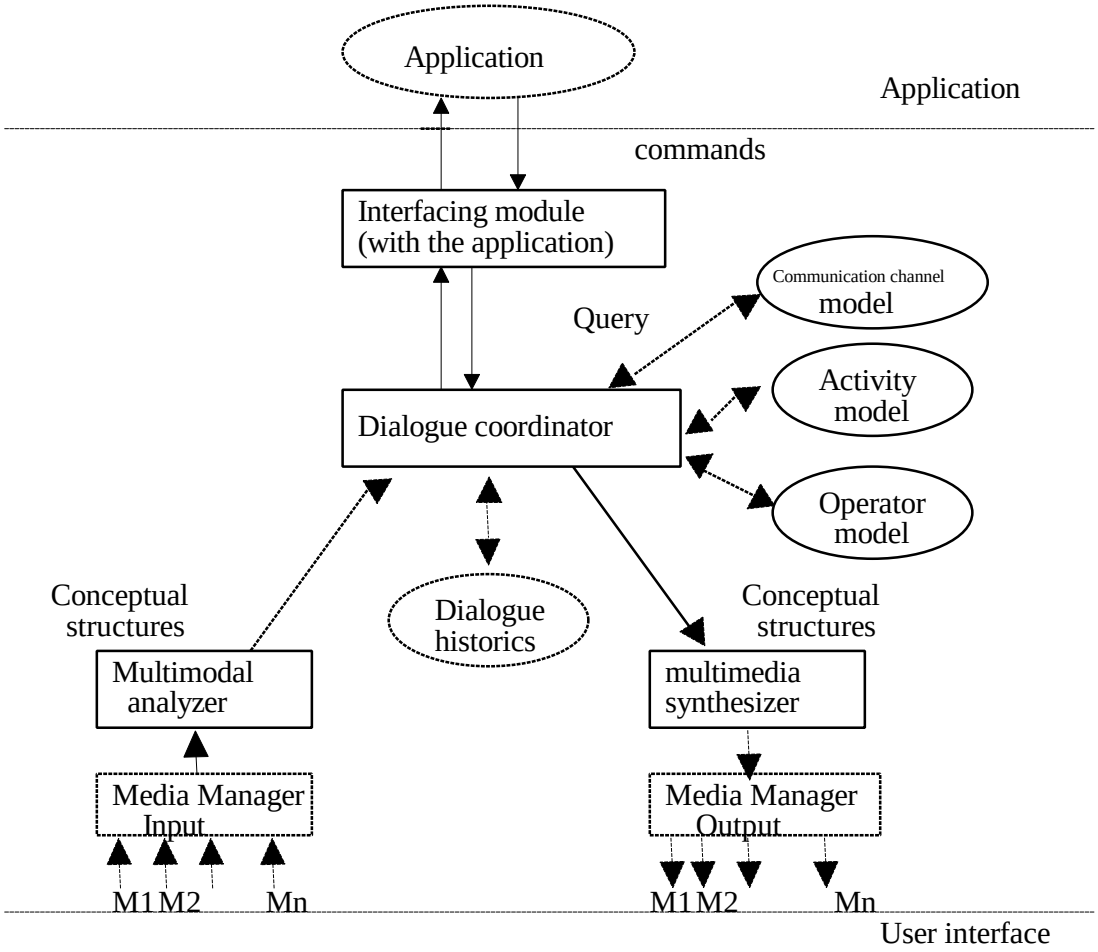


Figure 9 – L'architecture générale de MELODIA

- Le *module d'interface* (avec l'application), reçoit la description fonctionnelle de la requête de l'utilisateur qui est transmise par le coordinateur du dialogue et traduit la requête en formes interprétables par l'application. La traduction est bidirectionnelle (coordinateur du dialogue /application et application/coordinateur du dialogue).
- Le *multimedia synthesizer* reçoit une description fonctionnelle qui représente un message envoyé à l'opérateur. Le synthétiseur multimédia traduit le message et le présente à l'utilisateur.

L'architecture générique et modulaire de MELODIA le rend opérationnel avec différentes applications. Cette architecture est extensible au modèle «blackboard» et elle est basée sur des modules experts indépendants, selon les différentes opérations associées à chaque niveau. Le coordinateur du dialogue est partagé et il prend le rôle du superviseur dans l'architecture du «blackboard» [15].

2-4- MUNIX (multimodal UNIX)

MUNIX [46] est une application qui utilise la parole dans l'interaction entre l'humain et le système UNIX. Les ressources du système UNIX sont représentées comme des icônes dans un environnement «multi-window». Ces icônes représentent une grande variété d'objets comme des documents, des fichiers, des usagers etc. Les fenêtres peuvent être activées par une commande vocale et la modalité graphique est utilisée en association avec la parole pour désigner les ressources nécessaires pour une commande spécifique. Le modèle final de l'interaction adopté pour une commande UNIX présente trois entités du domaine UNIX (sans ordre imposé) :

Action - Objet(s) - Option(s)

Les objets des commandes sont représentés en utilisant le mode graphique (sélectionner par la souris). L'action et le choix des options sont donnés soit par la modalité parole ou bien par la modalité graphique. La grammaire définie dans MUNIX peut faire la distinction entre trois genres d'énoncés : prononcés par l'utilisateur, reconnus par la carte de reconnaissance et déduits de la grammaire de l'application.
L'architecture de l'interface multimodale de MUNIX est présentée dans la figure 10.

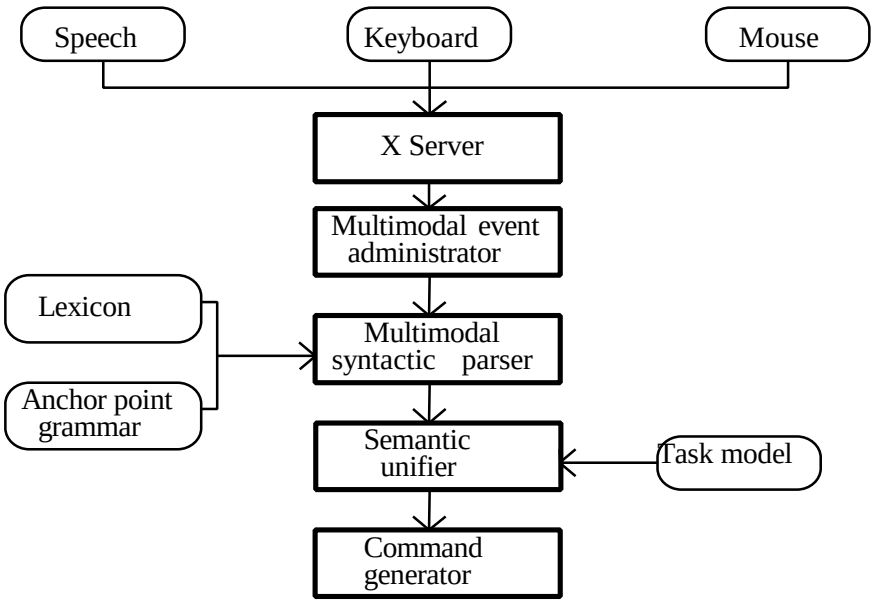


Figure 10 – L'architecture de MUNIX

2-5- PARTNER

PARTNER [52] est un système générique qui peut fournir un dialogue reconfigurable. PARTNER gère le dialogue et le traitement des entrées sorties.
Le système est implémenté sous forme d'un serveur auquel des clients peuvent se connecter. L'architecture sépare nettement le dialogue et l'application. Les composants de PARTNER sont :
1) l'input/output manager, 2) l'application manager et, 3) le dialogue manager (figure 11).

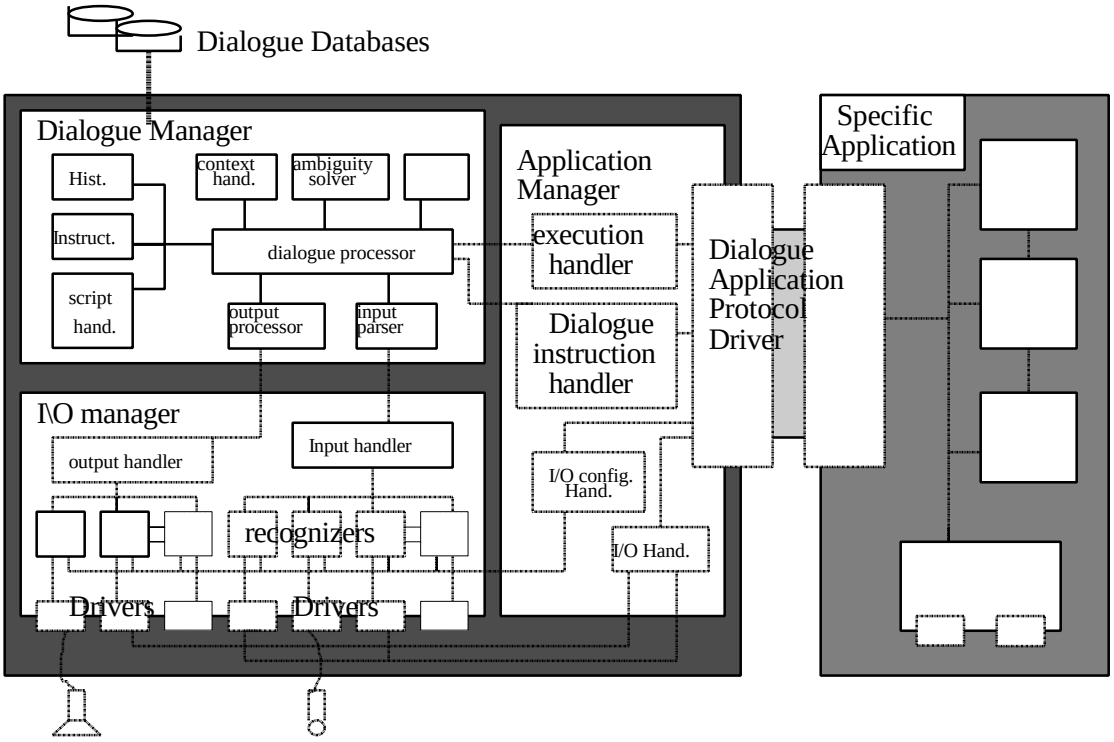


Figure 11 - Generic architecture of PARTNER

Dans PARTNER plusieurs mécanismes génériques et complémentaires d'assistance ont été introduits de manière à garder le dialogue cohérent et convivial. L'assistance est activée si besoin par l'utilisateur (il peut spécifier le degré d'assistance qu'il souhaite) ou automatiquement, par le serveur de dialogue lorsque des situations problématiques sont détectées. Les cinq mécanismes d'assistance de base sont les suivants :

1- *Dynamic proposal*
Ce mécanisme propose une réponse à l'utilisateur lorsque celui-ci pose une question de clarification.

2- *Automatic assistance*

Lorsque ce mécanisme est activé, plusieurs réponses sont proposées selon l'évolution de la tâche.

3- *Contextual help*

Les utilisateurs peuvent entrer en mode d'aide pour obtenir une liste de choix possibles à chaque tour de parole. Dans le cas de requêtes du système une réponse par oui ou par non est possible.

4- *Explanation*

Les utilisateurs peuvent demander des explications sur certaines meta-commandes.

5- *Language completion*

Ce mécanisme permet à l'utilisateur de savoir à chaque instant ce qu'il a le droit de dire et comment le dire.

Finalement PARTNER fournit un dialogue coopératif fondé sur des mécanismes d'aide conviviaux et hiérarchisés.

Section 1
Annexe 2

Bibliographie de la section 1

[1] Baecker R.M. & al. (1995). Speech, language, and audition, Baecker R.M. & al. (éds). Human computer interaction, Toward the year 2000, second edition, Morgan Kaufmann, San Francisco, p525-537

[2] Baudel T. (1991). Spécificités de l'interaction gestuelle dans un environnement multimodal, IHM'91, p11-16

[3] Baudel T. ; Beaudouin-Lafon M. (1993). CHARADE, Remote Control of Objects using Free-Hand Gestures, Communications of the ACM, vol 36, n°. 7, p28-35

[4] Bell D. ; Johnson P. (1994). General models of multimedia interaction, ERCIM, Multimodal human-computer interaction, p25-39

[5] Bellik Y. (1995). Interfaces multimodales, concepts, modèles et architectures, thèse, Orsay

[6] Bellik Y. ; Teil D. (1992). Les types de modalités, IHM'92 p22-28

[7] Bellik Y. ; Teil D. (1993). Specimen : un outil pour la spécification des interfaces multimodales, IHM'93, p9-16

[8] Bernsen N. O. (1995). A toolbox of output modalities ; Representing output information in multimodal interfaces, Bernsen N. O. ; Jensager F. ; Lu S. ; Verjans S. (Eds). Modality theory and information mapping, projet AMODEUS, rapport D15

[9] Bernsen N.O. (1994a). Foundations of multimodal representations, a taxonomy of representational modalities, Interacting with computers, vol 6, p347-371

[10] Bernsen N.O. (1994b). Why are Analogue Graphics and Natural Language both Needed in HCI?, Paterno F. (éd.) Design, specification and verification of interactive systems, Eurographics Workshop, Italie, p.165-179

[11] Bernsen N.O. (1994c). Modality theory : supporting multimodal interfaces design, Ercim'94, p13-23

[12] Bétrancourt M. (1992). Interaction texte/figure : effet de leur disposition spatiale relative sur l'apprentissage, rapport INRIA, n°1781

[13] Bisseret A ; Montarnal C. (1993). Stratégies de linéarisation lors de descriptions textuelles de configurations spatiales, rapport INRIA, n°1927

[14] Bisseret A. ; Spérandio J.C. (1968). Facteurs humains dans l'étude des dispositifs d'entrée d'information, Bulletin de CERP, Vol XVII n°4, p269-294

[15] [Bisson 92] : Bisson P. & Nogier J.F., *Interaction Homme-Machime multimodale : le système MELODIA*. Erg. IA'92, Biarritz, October, p.69-90.

[16] Bos E. ; Huls C. ; Claassen W. (1994). EDWARD : full integration of language and action in a multimodal user interface, International Journal of Human-Computer Studies, 40 (3), p473-496

[17] Bourguet M.L. (1992). Conception et réalisation d'une interface de dialogue personne-machine multimodale, thèse, INP, Grenoble

[18] Bradford J.H. (1995). The human factors of speech-based interfaces, SIGCHI bulletin, vol 27, n°2, p61-67

[19] Buxton W. (1991). The three mirrors of interaction : a holistic approach to user interface, Friend21'91, International symposum on next generation human interface, Tokyo, Japan, p25-27

[20] Buxton W. (1995). Speech, Language ; Audition. Chapter 8, Baecker R.M. & al. (Eds.) Readings in Human Computer Interaction, Toward the Year 2000, San Francisco, Morgan Kaufmann Publishers

[21] Cadoz C. (1992). Interface de communication instrumentale : clavier retroactif modulaire, l'interface des mondes réels et virtuels, Montpellier, p43-47

[22] Catinis L. ; Caelen J. (1994). Multimodal Human-Computer Interface, First Al-SHAM International Conference on Information Technology, Damascus - Syria , May 1994. p423-440

[23] Caelen J. ; Coutaz J. (1991). Interaction Homme-Machine multimodale, problèmes généraux, IHM'91, Dourdan, p41-58

[24] Caelen J.(1992), Compte rendu du workshop IHMM organisé par le GDR-PRC "Communication Homme-Machine", Dourdan 13-14 April 1992.

[25] Caelen J.(1993a). Interfaces multimodales, concepts et principes, article soumis à TSI, 1993.

[26] Caelen J. (1993b). Speech and multimodal interface, the case of ICPdraw, Workshop of ESCA'93

[27] Caelen J.(1994). Multimodal Human-Computer Interface, First Al-SHAM International Conference on Information Technology, Damascus - Syria , May 1994. p397-422

[28] Caldwell B.S. ; Paradkar P. (1995). Factors affecting user tolerance for voice mail message. International Journal of Human-Computer Interaction, Vol 7 n°3, p235-248

[29] Carbonell N. & al. (1994b). Etude empirique : usage du geste et de la parole en situation de communication homme-machine, ERGO'IA'94, p128-139

[30] Caro S. (1993). Eléments pour une typologie des unités textuelles de textes techniques, rapport INRIA, n°1930

[31] Chapelier L. (1996). Dialogue d'assistance dans une interface homme-machine multimodale, thèse, CRIN, Nancy

[32] Cohen P. R. (1992). The role of natural language in a multimodal interface, UIST'92, p143-149

[33] Condom J.M. (1992). Un système de dialogue multimodal pour la communication avec un robot manipulateur, thèse, Toulouse

[34] Conversy S. ; Beaudoin-Lafon M. (1996). Le son dans les applications interactives, Nouvelles Interfaces Homme-Machine, Jean Caelen (Ed)., Observatoire Francais des Techniques Avancées, Collection ARAGO, p65-81

[35] Coutaz J. & al. (1995). Four easy pieces for assessing the usability of multimodal interaction : the CARE properties, Interact'95, Lillehammer, Norway

[36] Coutaz J. (1992). Multimedia and multimodal user interfaces : a taxonomy for software engineering research issues, EWCHI, St Petersburg

[37] Coutaz J. ; Caelen J. (1991). A taxonomy for multimedia and multimodal user interfaces, ERCIM, p143-148

[38] Faure C. ; Arnold M. (1993a). L'interaction du point de vue des principes d'économie, IHM'93, p3-8

[39] Faure C. ; Julia L (1993b). Interaction homme-machine par la parole et le geste pour l'édition de documents :TAPAGE, Interface des modes réels et virtuels, p171- 180

[40] Frese M. & al. (1990). The effects of task structure and social support on user's errors and error handling, Interact'90, p35-41

[41] Gendrot H. (1994). Les interfaces multimodales, Exposé du DEA d'informatique communication homme-machine & ingéniérie éducative, Université du Maine

[42] Greenstein J.S. ; Arnaut L.Y. (1988). Ch 22, Input devices, Helander M. (Ed). Handbook of human-Computer interaction, Elsevier science publishers, North Holland

[43] Hauptmann A.G. ; McAvinney P. (1993). Gestures whith speech for graphic manipulation, International journal of man-machine studies, p231-249

[44] IHM'93 Synthèse de l'atelier\ Système d'analyse des interactions homme-ordinateur, "interface multimodaux, sous-groupe : formes de multimodalité en situation d'interaction utilisateur-machine"

[45] Julia L. (1995). Interface Homme-Machine multimodale pour la conception et l'édition de documents graphiques, Thèse, ENST Paris.

[46] Lefebvre p.(1993), Duncan G. & Poirier F., Speaking with computers: a multimodal approach, Eurospeech'93, 3rd European Conference on Speech Communication and Technology, Berlin, Germany 21-23 September 1993.

[47] MacKenzie S. (1995). Input devices and interaction techniques for advanced computing, Furness T.A. ; Barfield W. (Eds). Virtual environment and advanced interface design, New-York, Oxford university press, p 437-472

[48] MacKenzie S. ; Sellen A. ; Buxton W. (1991). A comparison of input devices in elemental pointing and dragging tasks, CHI'91, New-York, ACM, p161-166

[49] Martin J.C. (1995). Coopérations entre modalités et liage par synchronie dans les interfaces multimodales, Thèse, ENST, Paris95

[50] Mignot C. (1995). Usage de la parole et du geste dans les interfaces multimodales, étude expérimentale et modélisation, thèse, Nancy

[51] Mills Z. ; Prime M. (1990). Are all menus the same ? an empirical study, Interact'90, p423-427

[52] Morin P. & Junqua J.C.(1993), *Habitable interaction in goal-oriented multimodal dialogue systems*. Eurospeech'93, 3rd European Conference on Speech Communication and Technology, Berlin, Germany 21-23 September 1993.

[53] Néel F. (1994a). Communication orale avec la machine, récents développements, Technologie, santé, n°16, p87-96

[54] Néel F. (1994b). Reconnaissance de la parole, applications et perspectives, Compte rendu de l'OFTA 8-09-94, thème nouvelles interfaces Homme-Machine

[55] Nigay L. (1994). Conception et modélisation logicielles des systèmes interactifs : application aux interfaces multimodales, Thèse, Grenoble

[56] Nigay L. ; Coutaz J. (1993). A design space for multimodal system : concurent processing and data fusion. Interchi'92, p172-178

[57] Pleczon P. ; Chalard S. ; Saint Blancard M. (1994). Interface Homme-Machine multimodale pour un copilote intelligent d'aide à la conduite, ErgoIA'94, p108-118

[58] Stephanidis C. & Akoumianakis D.(1993), User modeling and multimodal representations: An approach combining alternative knowledge sources,

Workshop ERCIM on Multimodal Human-Computer Interaction, Nancy 2-4 November 1993.

[59] Wilson M.D., Sedlock D., Binot J-L. & Falzon P.(1991), An architecture for multimodal dialogue, ETRW, Second Venaco Workshop, The structure of multimodal dialogue, Acquafredda di Maretea - Italy September 16-20 1991.

[60] Zanello M.L. (1993). Observation des pratiques exploratoires comme outil de test de logiciel, mémoire de DEA de psychologie sociale, Grenoble

[61] Zanello M.L. ; Caelen J. ; Bisseret A. (1996). Une approche centrée tâche de la multimodalité, IHM'96, p99-104

[62] Zanello M.L. (1997), L'utilisateur et l'interface multimodale : contribution à la connaissance sur son utilisation et sa gestion . Thèse Université Joseph Fourier, Grenoble.

Section 2

L'usage de la multimodalité

Chapitre 3 : L'usage de la multimodalité : une étude expérimentale

Chapitre 4 : Comportement de l'usager avec un assistant

Chapitre 5 : Comportement de l'usager avec une interface multimodale

Chapitre 6 : Conclusions sur l'étude expérimentale sur l'usage de la multimodalité

Annexe 1 : Bibliographie de la section 2

Chapitre 3

L'usage de la multimodalité : une étude expérimentale

L'usage de plusieurs modes et/ou modalités en communication homme-machine est conditionné par l'efficacité et l'utilisabilité des systèmes interactifs, et reste centré sur l'intérêt que peuvent apporter les interfaces multimodales. Il serait vain de concevoir de telles interfaces sans un espoir d'un usage effectif même si cet usage passe à terme par une phase d'apprentissage. La question est donc de prédire de manière aussi précise que possible, la portée d'un tel usage sans bien sûr recourir à des expérimentations en vraie grandeur qui nécessiteraient la réalisation de l'interface elle-même. Pour répondre à cette question, différentes expériences ont été réalisées [12], [21], [20], [22] en utilisant la technique dite *Magicien d'Oz (Moz)*, technique mise au point dans d'autres secteurs d'utilisation (notamment par Richard, Morel, Falzon, Amalberti, etc.). Carbonell et al. [10] ont étudié le comportement multimodal des usagers par rapport à une grille de critères, à savoir : exclusif, concurrent, alterné et synergique [7]. Les résultats semblent mitigés quant à l'usage effectif de la multimodalité — en particulier il ne semble pas évident que l'usage synergique des modes soit aussi fréquent qu'on pouvait le penser a priori. A la suite de tels résultats, on est en face de deux réactions possibles : (a) remettre en cause la technique Moz elle-même (ce dont de nombreux auteurs ne se privent pas de faire) et/ou (b) choisir une autre grille de critères pour poursuivre les expériences :

(a) remettre en cause une technique sans proposer d'alternative reste stérile. Il nous semble préférable de considérer les résultats Moz pour ce qu'ils sont et de les replacer dans un cadre comparatif plus vaste, en variant les approches et méthodes de mesure, les consignes et les contraintes des expériences,

(b) il faut remarquer que la grille {exclusif, concurrent, alterné et synergique} décrit un système multimodal et non le comportement de l'usager. Depuis les travaux cités ci-dessus, différents chercheurs ont apporté des contributions nouvelles à une typologie des interactions multimodales du point de vue, de l'usager : les propriétés CARE [14], [18], [23] et les propriétés étendues T-CCARE [24].

Dans le chapitre 4 et avec ces points de vue (a) et (b), nous proposons une nouvelle expérimentation, à savoir évaluer la multimodalité sous l'angle des propriétés CARE et en nous plaçant en situation de communication

humaine contrainte par un support informatique dans le cadre d'une tâche donnée : il s'agit pour un usager de réaliser un dessin avec l'aide d'un assistant, tous les deux se trouvant côte à côte et face à un écran d'ordinateur (il n'y a plus d'ambiguïté pour le sujet sur la présence ou non d'un compère). Le choix de ce type de tâche et de situation sont justifiés par le souci de continuer et d'approfondir une démarche antérieure [21] mais surtout de progresser de façon maîtrisée pour obtenir des résultats ultérieurement comparables.

Dans le chapitre 5, et dans le but de progresser dans l'étude de l'usage de la multimodalité, nous proposons une expérimentation similaire à celle du chapitre 4, mais dans ce cas là, l'usager humain doit communiquer avec la machine sans avoir un assistant humain mais à l'aide d'un logiciel multimodal.
Dans le cadre de cette expérience on se trouve dans une situation d'interaction homme-machine réelle. Les problèmes liés au contact direct avec la machine seront cités et leurs impacts sur les résultats seront précisés en faisant la comparaison avec les résultats de l'expérience précédente.

Les deux chapitres suivants présentent les deux versions de l'expérience et le chapitre 6 fait une conclusion générale sur notre travail concernant l'usage de la multimodalité.

L'étude est basée sur les propriétés CARE qui sont présentées dans le chapitre 2.

Chapitre 4

Comportement de l'usager avec un assistant

1- Introduction

Ce chapitre décrit une expérience dans laquelle un groupe de sujets effectue une tâche de dessin avec un logiciel approprié, sous l'assistance d'un expert. L'objectif de l'expérience est d'étudier le comportement multimodal des sujets sous l'angle des propriétés CARE (Complémentarité, Assignation, Redondance, Equivalence) [14].

2- Description et conditions de l'expérience

L'objectif de la tâche proposée aux sujets est de dessiner quatre figures (fig. 1) dans un environnement de dessin "Draw" sur PC à partir d'éléments géométriques simples (triangles, carrés, cercles, lignes verticales et horizontales) (fig. 2) rangés dans une palette placée sous la zone de travail, vide au départ.

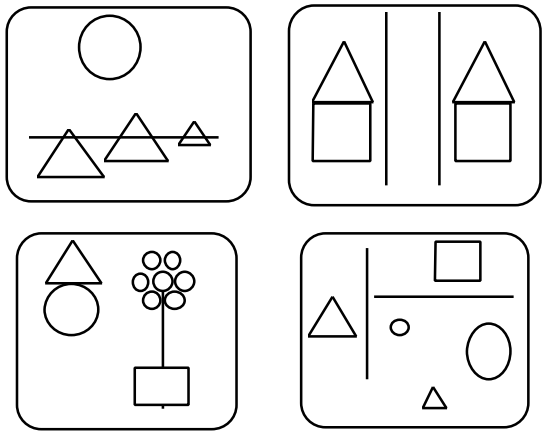


figure 1 : les 4 figures à dessiner

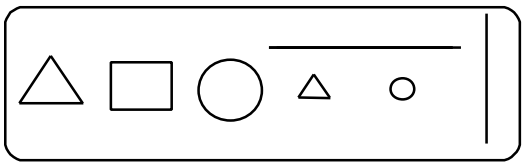


figure 2 : la palette des objets disponibles

Les scènes figuratives sont de complexité variable vis-à-vis de la symétrie géométrique, de la symbolique de la scène, de la répétitivité des éléments et de la forme. Nous avons volontairement limité le nombre de figures pour épargner la fatigue des sujets tout en essayant de faire varier au maximum ces paramètres de complexité d'une figure à la suivante (sans atteindre toutefois une combinatoire complète).

Afin de faire émerger des comportements multimodaux récurrents nous avons imposé aux sujets un ordonnancement des actions : nous comptions ainsi régulariser la charge de planification des actions pour tous les sujets et réduire son influence sur le choix des modes. Sans prétendre avoir effacé toute influence, cela nous a permis tout au moins de situer les sujets dans un même cadre normatif de tâche. Au cours de la session, le sujet joue le rôle d'instructeur vis-à-vis de son partenaire qui a un rôle d'assistant. Cet assistant est toujours le même pour tous les sujets ; il a subi un entraînement précis de façon à se comporter d'une manière aussi reproductible que possible. Nous avons encore contraint l'expérience en imposant au sujet un vocabulaire de commande limité et une syntaxe restreinte non ambiguë afin que l'assistant n'ait aucune difficulté ni marge d'interprétation d'une part et que d'autre part les sujets soient tous placés dans les mêmes conditions langagières ne favorisant pas *a priori* l'usage de la multimodalité. Cette nécessité s'imposait par ailleurs pour ne pas privilégier un mode par rapport à un autre c'est-à-dire pour annuler l'effet de préférence d'un mode d'expression — éventuellement plus riche — au détriment d'un autre : c'est pourquoi nous avons appauvri le mode langagier (notons au passage que ceci correspond mieux également aux performances actuelles des systèmes en reconnaissance automatique de la parole). Finalement, le sujet peut :

• énoncer des ordres multimodaux (parole et geste de désignation avec le doigt sur l'écran) ou monomodaux (parole seule) à l'adresse de l'assistant qui exécute ces ordres,

• dessiner lui-même la figure (ou un élément de la figure) avec les possibilités (souris + clavier) que lui offre le logiciel Draw.

Ceci définit les trois modes suivants :

(pdd) parole et désignation du doigt sur l'écran,
(p) parole seule,
(s) souris + clavier (sans parole).

Remarquons que la modalité gestuelle dans le mode (pdd) ne passe pas par un capteur. Nous nous attendons donc à ce que ce mode soit plus spontané et plus "naturel" qu'un geste médiatisé par la souris. Par défaut, comme il est moins précis, il engendre un coût de traitement inférentiel de la cible plus grand de la part de l'interprète et donc peut-être une demande implicite de précision langagière plus forte de sa part vis-à-vis du sujet (pour le moment l'assistant humain mais plus tard la machine). Ces deux faits font que peut-être ce mode mixte (pdd) sera plus propre à supporter la multimodalité que d'autres modes mixtes étudiés jusque là, fondés sur la souris.

Pour rendre la situation plus engageante et plus réaliste pour le sujet nous lui avons demandé d'exécuter le dessin en un minimum de temps possible (mais cette contrainte a peut-être engendré un biais vis-à-vis de l'assistant que certains pouvaient considérer comme un examinateur).

L'assistant avait pour rôle d'interpréter les ordres du sujet et de les exécuter. Il avait pour consigne de ne pas exécuter les ordres qui contenaient des mots ne faisant pas partie du vocabulaire ou des énoncés ayant une syntaxe trop complexe (propositions relatives, reprises, etc. étaient interdites). Le sujet et l'assistant avaient sous les yeux en permanence la liste des mots autorisés afin de ne pas pénaliser l'usage éventuel de la langue par rapport aux autres modes par un effort mémoriel supérieur. La sémantique devait se réduire impérativement au schéma ***Action - Objet - Lieu*** sous peine d'être rejetée par l'assistant, avec :
Action = [dessiner] [déplacer] [effacer] [stop]
Objet = [ligne verticale] [ligne horizontale] [grand triangle] [petit triangle] [grand cercle] [petit cercle] [carré] [*anaphoriques* = il, elle, le, la, les, etc.] [*déictiques* = ce, cette, ça, celui-ci, celui-là, etc.]
Lieu = [PositionRelative(objet)] [Position] [PositionAutoRelative]
PositionRelative(objet) = [au-dessus-de(objet)] [au-dessous-de(objet)] [à-gauche-de(objet)] [à-droite-de(objet)]
Position = [ici] [là] [au-centre] [en-haut] [en-bas] [à-droite] [à-gauche]
PositionAutoRelative = [plus haut] [plus bas] [plus à gauche] [plus à droite]

Exemples : « dessiner grand cercle au-dessus de la ligne verticale »,
« déplacer grand triangle à droite » - « stop »,
« déplacer grand cercle plus à droite ».

L'action dessiner un objet s'effectue en déplaçant un objet de la palette et en le posant sur l'espace de dessin sur l'écran. Effacer un objet s'effectue en désignant l'objet et en cliquant sur la touche « delete ».
Malgré des contraintes aussi strictes, nous n'étions pas encore totalement assurés que l'assistant (pourtant entraîné) les respecte scrupuleusement et sans fatigue. Pour atténuer ce biais possible nous avons réduit la durée des expériences au minimum et conservé le même assistant pour toutes les expériences (ainsi tous les sujets étaient en face des mêmes conditions interlocutoires).

3- Déroulement de l'expérience

Les sujets n'étaient pas tous familiers des logiciels de dessin. Sur les 26 sujets testés 3 se trouvaient pour la première fois devant un écran et 5

étaient habitués à utiliser l'informatique. Les autres étaient de niveau variable (occasionnels). Les sujets novices pouvaient se familiariser avec le logiciel avant l'expérience pour se décontracter et éviter des blocages psychologiques (en effet il pouvait s'instaurer involontairement des relations de compétence hiérarchique avec l'assistant). Les domaines d'activité des sujets étaient très larges : des programmeurs, des ingénieurs de plusieurs spécialités, des médecins, des économistes, des secrétaires, des étudiants, des économistes.

Les expériences étaient filmées et analysées manuellement (nous avons utilisé une camera vidéo pour filmer toutes les expériences et nous les avons analysées ultérieurement sur un écran de télévision) : on a relevé pour chaque action les propriétés CARE des sujets. Les actions retenues par l'analyse sont données dans l'arbre (fig. 3), c'est-à-dire essentiellement les actions à résultat immédiat.

Les analyses s'intéressaient aussi aux relations temporelles dans le cas des commandes avec déictiques [là] [ici] [ça] accompagnées de gestes de désignation du doigt sur l'écran. Ces relations temporelles ont été notées (k>D, k=D, k<D, k<<D) pour indiquer respectivement que (a) le geste est consécutif au mot prononcé, (b) le geste est sensiblement simultané avec le mot prononcé, (c) le geste précède le mot prononcé, (d) le geste précède la séquence parlée qui contient la deixis et le doigt reste posé sur l'écran ensuite pendant l'énoncé verbal.

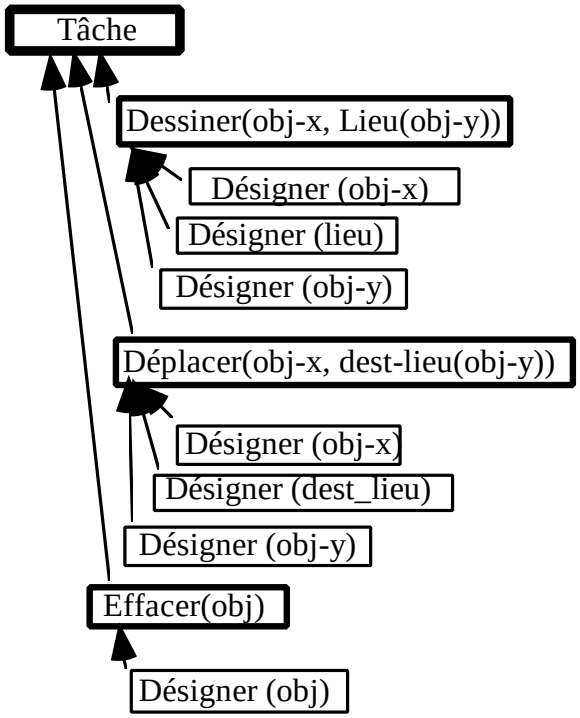


figure 3 : Dans les cadres gras, les actions multimodales retenues dans l'analyse dans l'ordre de leur complexité sémantique décroissante (avec leur décomposition en actions plus élémentaires).

Enfin, toutes les observations intéressantes étaient notées, ainsi que les commentaires des sujets et leurs réactions en cas d'erreur d'exécution de l'assistant ou de leurs propres erreurs.

4- Résultats tirés de l'expérience

4-1- L'équivalence des modes

L'équivalence des modes désigne le fait d'obtenir un même résultat en utilisant différentes modalités [17]. On peut distinguer deux notions : l'équivalence de résultat et l'équivalence fonctionnelle.

équivalence de résultat : on peut constater que les modes proposés dans cette expérience (pdd) (p) et (s) sont équivalents vis-à-vis du résultat final (obtenir une figure). En effet chaque mode est suffisant à lui seul pour effectuer la tâche à accomplir. L'analyse des statistiques montre ce résultat : sur les 104 figures dessinées par les 26 sujets, 40% ont été dessinées de bout en bout avec le mode (pdd), 6% avec le mode (p), 16% avec le mode (s), le reste des figures ayant été dessiné en variant les modes au cours de la séance..

équivalence fonctionnelle : L'équivalence fonctionnelle étant basée sur la précision et l'aisance d'usage de chaque mode [17], nous concluons de notre expérience qu'il n'y a pas d'équivalence fonctionnelle entre les modes. En effet nous avons considéré pour le paramètre *précision* les actions Désigner(objet) et Désigner(lieu) — qui seules nécessitent une certaine précision d'exécution — et nous avons relevé le nombre de fois où l'un et/ou l'autre mode ont été utilisés. Les résultats sont les suivants :

toutes catégories de sujets (26 sujets)

(pdd)	(p)	(s)
66%	14%	20%

tableau 1 : usage des modes

catégorie expert (5 sujets)

(pdd)	(p)	(s)
81%	9%	10%

tableau 2 : usage des modes

catégorie occasionnel (18 sujets)

(pdd)	(p)	(s)
53%	17%	30%

tableau 3 : usage des modes

catégorie novice (3 sujets)

(pdd)	(p)	(s)
78%	22%	0%

tableau 4 : usage des modes

On constate, à partir de ces 4 tableaux (tableau 1 à tableau 4), qu'il y a en moyenne et dans chaque catégorie un usage prédominant du mode pdd. L'usage des deux autres modes (p et s) dépend de la catégorie d'usager, les plus réticents à la souris étant les novices, et les plus favorables les occasionnels.

Le tableau 5 affine quelque peu ces résultats : il présente les fréquences d'utilisation du geste de désignation du doigt et de la désignation verbale des lieux et des objets pour les trois catégories de sujets.

Expert			
désignation-lieu		désignation-objet	
doigt	parole	doigt	parole
90%	10%	22%	78%
Occasionnel			
désignation-lieu		désignation-objet	
doigt	parole	doigt	parole
87%	13%	23%	77%
Novice			
désignation-lieu		désignation-objet	
doigt	parole	doigt	parole
88%	12%	22%	78%

tableau 5 : usage de la modalité "doigt" ou "parole" pour désigner un lieu ou un objet selon le type d'usager

Ces résultats (remarquablement stables) montrent que le lieu est désigné préférentiellement par un geste, tandis qu'un objet est plutôt désigné verbalement par son nom ou par un de ses traits pertinents (de couleur, de position, etc., par exemple "déplace le rouge à gauche"), ceci indépendamment de la catégorie des sujets.

Le paramètre *aisance d'usage* est plus difficile à mesurer par observation directe, il aurait fallu installer un dispositif de mesure du temps d'exécution des actions dans le logiciel et surtout disposer d'une interface réellement multimodale. Nous en sommes donc restés à une estimation subjective obtenue à la suite des commentaires donnés par les sujets eux-mêmes : 22/26 des sujets ont préféré le mode (pdd).

Les résultats sont donc clairs dans notre expérience : le mode (pdd) (parole et désignation du doigt sans contrainte instrumentale) multimodal par essence, qui semblait *a priori* le plus "naturel" emporte effectivement l'adhésion des sujets qui l'utilisent dans la majorité des cas.

4-2- L'assignation des modes

Pour estimer le degré d'adéquation d'un mode à un type d'action (assignation) nous avons calculé pour chaque action et selon son degré de réalisation (succès ou échec), le pourcentage relatif des modes utilisés. Avec plus de détails par types d'action nous avons :

(a) pour les cas de réussite de l'action

"dessiner"		
(s)	(p)	(pdd)
18,6%	6,6%	74,8%
"déplacer"		
(s)	(p)	(pdd)
5,8%	31,4%	62,8%
"effacer"		
(s)	(p)	(pdd)
22%	27%	51%

(b) pour les cas d'échec de l'action (les valeurs indiquées sont les nombres d'actions échouées qui sont en elles-mêmes peu nombreuses par rapport au nombre total d'actions tentées)

"dessiner"		
(s)	(p)	(pdd)
1	1	15
"déplacer"		
(s)	(p)	(pdd)
0	2	3
"effacer"		
(s)	(p)	(pdd)
0	2	0

tableau 6 : usage des modes selon le type d'action

Remarquons tout d'abord que l'action "déplacer" est assez souvent effectuée en préambule de l'action "dessiner" c'est-à-dire que le sujet commence à placer un objet sur l'écran à partir de la palette, puis utilise l'action "dessiner" pour positionner plus finement l'objet. Pour ce faire il utilise plus volontiers la souris qui permet une désignation plus précise du lieu. Cela explique pourquoi la souris est davantage sollicitée dans "dessiner" que dans "déplacer" (de plus pour une désignation vague la parole est plus économique). Une assignation se dessine donc entre l'usage de la parole (p) et l'usage de la souris (s). Pour l'action "effacer", parole et souris s'équilibrent : des raisons de meilleure "sécurité" n'ont pas fait préférer (s) contre (p). Remarquons que le mode (pdd) reste en tête des usages et ceci malgré les échecs plus nombreux (mais encore rares donc probablement supportables) — les échecs qui ont été répertoriés ici ne sont dus qu'à des erreurs du sujet (et non celles qui proviennent d'une mauvaise interprétation de l'assistant).

4-3- La complémentarité

La complémentarité se définit pour une commande multimodale par rapport aux informations véhiculées par chaque modalité : lorsque ces informations se complètent pour donner un sens à l'énoncé multimodal, on dit que les modalités sont complémentaires — cela ne présage pas de leur usage synchrone (et donc synergique) ou asynchrone (alterné). Dans notre expérience cela ne concerne que le mode (pdd) : la complémentarité se mesure dans une tâche en comptant le nombre d'actions qui, utilisent des modalités complémentaires sur le total des commandes multimodales. Les résultats donnent pour la tâche complète de dessin et tous sujets confondus :

"dessiner	"déplacer"	"effacer"	moyenne pour toutes les actions
90.2%	59.3%	58%	66.6%

tableau 7 : usage de la complémentarité selon le type d'action

Ces résultats montrent que la complémentarité est davantage mise à profit par les usagers pour les opérations sémantiquement complexes (voir fig. 3). En effet le recours à la complémentarité dans un énoncé peut être interprété comme une recherche d'économie du côté de la production : il est évident qu'un énoncé dans lequel chaque élément est nécessaire et suffisant est plus "économique" qu'un énoncé redondant qui aurait toutes choses égales par ailleurs, les mêmes effets. C'est la raison pour laquelle une commande déjà coûteuse à formuler au plan sémantique aura des chances d'être moins redondante qu'une commande plus simple. Il est évident que ceci peut être contredit par le critère de "fiabilité" qui va à l'encontre de l'économie de production et qui favorise l'économie de résultat. Nous discutons ce point maintenant.

4-4- La redondance

La redondance multimodale à laquelle on s'intéresse ici (il peut en effet y avoir aussi une redondance monomodale) se définit en opposition avec la complémentarité : elle renvoie aux informations redondantes véhiculées par plusieurs modalités qui ont les mêmes résultats dans l'interprétation de la commande. Du point de vue de l'énonciateur, la redondance ne peut représenter une économie de production, mais du point de vue du destinataire cela peut représenter une économie d'interprétation : elle lui sert souvent à désambiguïser la commande ou à l'éclaircir. En retour, la redondance est un moyen pour l'énonciateur de fiabiliser sa commande. Donc, complémentarité et redondance s'équilibrent sur l'axe économie-fiabilité du point de vue de l'énonciateur.

Le tableau ci-après indique pour les actions “dessiner”, “déplacer” et “effacer” les redondances pour l’expression d’un lieu ou la référence d’un objet. On retrouve bien le fait que l’opération “à risque” (effacer) devant être la plus fiable est aussi celle qui est la plus redondante. Avec ce même type de raisonnement on peut avancer que l’objet de l’action “dessiner” est l’objet de plus d’attention de la part du sujet que le lieu.

“dessiner”		**“déplacer”**		**“effacer”**
désignation		désignation		désignation
objet	lieu	objet	lieu	objet
7.4%	2.4%	19.8%	20.9%	42%

tableau 8 : usage de la redondance selon le type d’action et les éléments sémantiques à l’intérieur des énoncés de ces actions

On a remarqué par ailleurs deux tendances chez les sujets qui ont passé une deuxième fois l’expérience (6 sujets): (1) voulant être plus précis dans la désignation des lieux (certainement pour augmenter la fiabilité de leurs commandes et donc l’efficacité générale de leur action) ils ont augmenté le taux de redondance des modalités (par exemple en disant “ici à droite du cercle” et en montrant la position correspondante). (2) En même temps, s’étant rendus compte que la désignation des objets n’était pas très ambiguë dans cette application (les objets sont facilement discriminables et nommables), ils ont maintenu le même taux de complémentarité pour la dénomination des objets[1].

5- Les relations temporelles dans les cas de deixis

Un second objectif de l'expérience était de mesurer les relations temporelles entre la prononciation de mots déictiques (ici, là, ça) et la désignation gestuelle des objets et des lieux correspondants. Les analyses ont montré que dans 67.3% des cas de deixis l'usager a désigné avec le doigt en même temps — cette notion de synchronie se rapporte à la perception d’un observateur humain — qu’il prononçait le déictique (la synchronie est notée : k=D) : nous appelons ce cas (a) synchronie vraie synergique. Deux autres cas de synergies ont été observées :

[1]Mais ne voulant pas étudier spécialement les phénomènes liés à l’apprentissage, nous n’avons pas poussé cette expérience avec de nombreux sujets, il faut donc prendre ces résultats avec précaution et les considérer comme des tendances purement qualitatives.

(b) D<[k]: asynchronie anticipative (ou fausse synchronie) de la parole sur le geste,

(c) [k]<D : asynchronie anticipative du geste sur la parole,

ainsi que deux cas d'asynchronie alternée :

(d) [k]<<D : le geste précède le temps de parole (et a fortiori la deixis contenue dans la séquence de parole),

(e) D<<[k] : le geste survient nettement après le temps de parole.

[k]<<D	[k]<D	[k]=D	D<[k] et D<<[k]
3.2%	28.8%	67.3%	0.7%

tableau 9 : taux de synchronie et d'asynchronie dans les actions multimodales (voir significations de D et [k] dans le texte).

On notera le faible nombre de cas où la parole anticipe sur le geste, cela est peut-être dû à l'usage du doigt qui "traîne" sur l'écran entre deux actions. Dans les commandes de type "mets ça là" les désignations gestuelles ont toujours respecté l'ordre des déictiques c'est-à-dire d'abord pour satisfaire "ça" et ensuite pour satisfaire "là".

6- Observations générales

Ellipses

Les formes d'expression comportaient naturellement des ellipses, anaphores et déictiques mais aussi des ellipses déictiques (omission de ça par exemple dans "déplace" + geste).

Réparation des erreurs

L'assistant a fait quelques erreurs dans l'exécution des commandes. Les erreurs ont porté sur la taille ou la position d'objets, ou sur la référence de l'objet, ou plus rarement sur la commande. Le sujet n'a pas toujours réparé ces erreurs, il a même parfois pensé que cette erreur venait de lui. Quand il a réparé les erreurs il a réitéré sa commande (seule stratégie qui lui était permise dans cette expérience) plus comme s'il s'agissait d'une mauvaise audition que d'une mauvaise compréhension de la part de l'assistant.

1-Réparation monomodale : Les améliorations de l'expression verbale n'ont pas été sensibles, en particulier les cas d'utilisation d'ellipses pour réparer les situations d'échec ont été de 67% ce qui

montre bien que le sujet ne tente pas d'améliorer la précision de son expression même dans une situation difficile, à moins qu'il ne veuille au contraire attirer l'attention que sur l'élément incriminé, comme dans l'exemple suivant :.
commande : "dessine un carré là"
exécution : un grand cercle est dessiné
réparation : "un carré ..... un carré ..."

2-Réparation multimodale : Le tableau 10 donne quelques éléments d'appréciation sur la manière dont les sujets réparent les erreurs en s'aidant de la multimodalité, soit en réitérant la commande sur un même mode soit en changeant de mode. Le mode pdd étant dans tous les cas renforcé, il semble qu'il apparaisse plus sécurisant pour le sujet.

	mode ayant provoqué l'erreur	
mode correction	p	pdd
s	0%	0%
p	12%	6%
pdd	88%	94%

tableau 10 : fréquence d'utilisation des modes dans la phase de réparation

Stratégies de placement des objets
On peut tirer quelques observations qualitatives sur le déroulement de la tâche :
- certains usagers procédaient par approximation successive pour placer les objets. Les cas d'ambiguïté de déictiques devaient alors être résolus par l'assistant de façon contextuelle.
- quelques usagers désignent les lieux ou les objets par entourage.
- pour déplacer un objet, certains usagers font re-dessiner un nouvel objet identique au premier par l'assistant puis lui font effacer l'ancien.
- quelques cas d'attente importante entre l'énoncé de la commande et la désignation gestuelle se sont produits : des usagers attendaient le "feed-back" de la machine avant de préciser le lieu de destination de l'objet.

7-Conclusion et prolongements

Les résultats de l'expérience donnent une bonne idée sur le comportement multimodal de l'utilisateur — dans la situation de communication

humaine contrainte dans laquelle nous l'avons placé et pour une tâche de dessin assisté.
Par rapport à des résultats d'autres auteurs ([10] par exemple), l'usage de la multimodalité semble plus important : cela est peut-être dû au fait qu'ici le geste n'est pas médiatisé par un capteur et que la communication se situe dans un contexte humain. Nos résultats s'apparentent à ceux de Bellik [2] concernant l'évaluation de l'interface LIMSIDraw réellement implémentée : cela permet d'avancer que des simulations peuvent être mises en place valablement pour prédire des résultats cohérents et que des expériences telles que celles-ci peuvent être poursuivies.
En ce qui concerne les propriétés CARE, nous avons observé une assignation naturelle des modes selon la tâche ainsi qu'une tendance à fiabiliser l'interaction par l'usage de la redondance (également dans [1]). Ce facteur prend moins d'importance si l'action est complexe ou à faible risque : il semble que des considérations d'économie l'emportent alors pour limiter l'usage de la redondance.

Dans le chapitre qui suit nous allons nous mettre dans une situation d'interaction homme-machine réelle et pour étudier les effets du contact direct avec la machine en nous mettant dans les mêmes conditions de tâche que dans cette expérience.

Chapitre 5

Comportement de l'usager avec une interface multimodale

1- Introduction

Ce chapitre décrit une expérience dans laquelle un groupe de sujets effectue une tâche de dessin avec un logiciel approprié équipé d'une interface multimodale. L'interface multimodale est capable de synchroniser les différentes modalités et les intégrer dans la construction du sens de la commande. L'objectif de l'expérience est d'étudier le comportement multimodal des sujets sous l'angle de propriétés CARE[14] et de les comparer avec l'expérience précédente.

2- Description et conditions de l'expérience

L'objectif de la tâche proposée aux sujets est de dessiner quatre figures (fig. 1) dans un environnement de dessin (logiciel spécialisé avec une interface multimodale « FIRAS ») sur PC à partir d'éléments géométriques simples (triangles, carrés, cercles, lignes verticales et horizontales) (fig. 2) rangés dans une palette placée au dessus de la zone de travail, vide au départ.

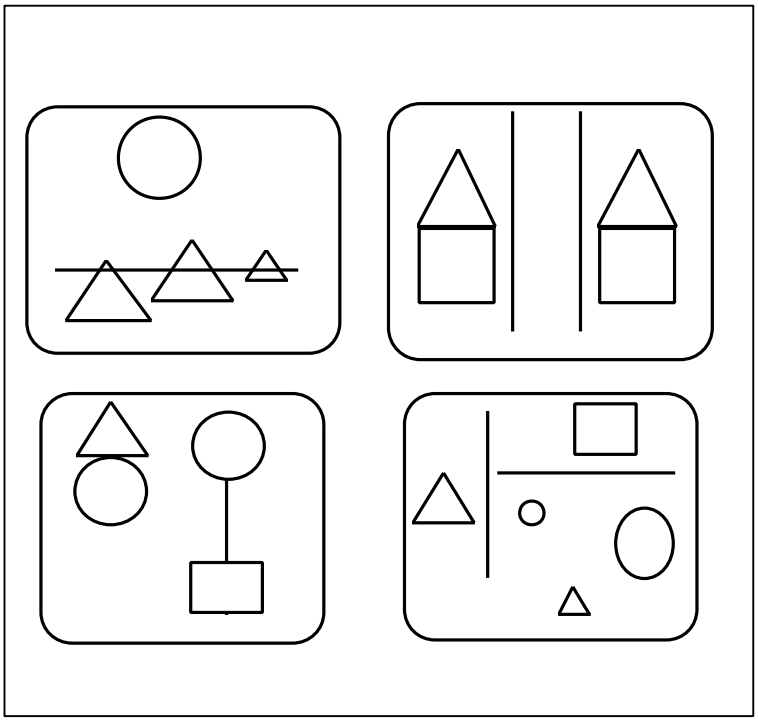


figure 1 : les 4 figures à dessiner

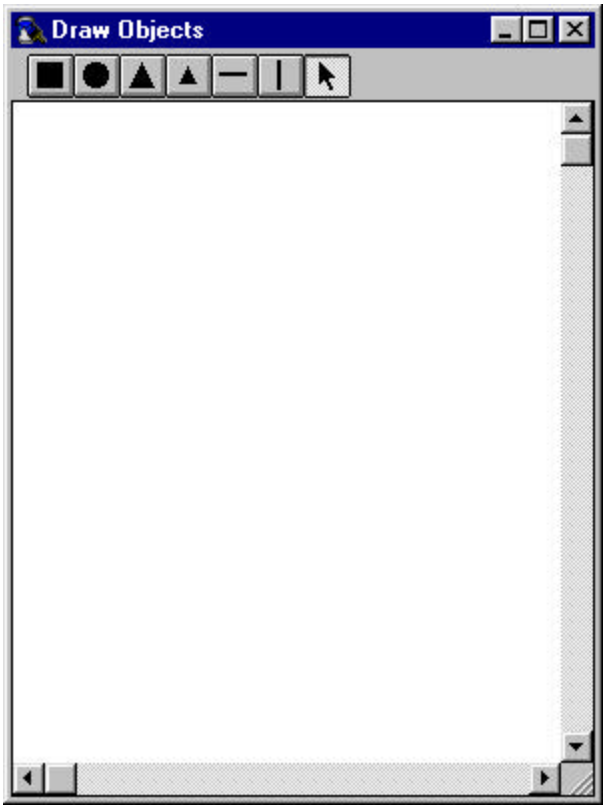


figure 2 : la palette des objets disponibles et la situation de départ

Les scènes figuratives sont les mêmes que celles de l'expérience précédente. L'ordonnancement des actions est imposé pour réguler la charge de planification des actions. Au cours de la session, le sujet à le choix d'accomplir la tâche en utilisant un ou plusieurs modes d'interaction. Le vocabulaire de commande est limité et la syntaxe est restreinte. Nous sommes donc dans des conditions identiques de celles de l'expérience précédente (avec assistant). Le sujet peut :

• énoncer des ordres multimodaux (parole et geste de désignation avec le bouton droit de la souris) ou monomodaux (parole seule),

• dessiner en manipulation directe la figure (ou un élément de la figure) avec les possibilités (souris + clavier) que lui offre le logiciel FIRAS.

Ceci définit les trois modes suivants :
(pdd) parole ou clavier et désignation par un click souris par le bouton droit de la souris,
(p) parole seule,
(s) souris + clavier (sans parole).

La sémantique devait se réduire impérativement au schéma ***Action - Objet - Lieu*** sous peine d'être rejetée par le logiciel FIRAS, avec :

Action = [dessiner] [déplacer] [effacer]
Objet = [ligne verticale] [ligne horizontale] [grand triangle] [petit triangle] [cercle] [carré] [*déictiques* = ça] [*anaphoriques* = il]
Lieu = [PositionRelative(objet)] [Position] [PositionAutoRelative]
PositionRelative(objet) = [au-dessus-de(objet)] [au-dessous-de(objet)] [à-gauche-de(objet)] [à-droite-de(objet)]
Position = [ici]
PositionAutoRelative = [plus haut] [plus bas] [plus à gauche] [plus à droite]

(Cette syntaxe est en fait en anglais – voir le détail du logiciel FIRAS)

3- Le Système :

Le système utilisé pour l'interaction avec l'utilisateur est un logiciel de dessin qui accepte des commandes écrites sur un prompt de commandes avec une syntaxe bien définie. Ces commandes peuvent être passées par la parole si on lance le logiciel de la reconnaissance de la parole ECHO-Lets'go ou (ECHOGO). Dans les deux paragraphes qui suivent nous présentons les deux logiciels (ECHO et FIRAS) en précisant la taxonomie de l'ensemble du système interactif.

3-1- Le logiciel ECHO :

ECHO est une application tournant sous Windows95, qui permet à un utilisateur de donner oralement des commandes à des applications Windows95. La reconnaissance vocale est effectuée à l'aide d'une bibliothèque de fonction développée à l'Institut de la Communication Parlée (ICP) puis au laboratoire CLIPS-IMAG, autorisant la reconnaissance de mots connectés. ECHO est basé sur le principe des chaînes de Markov et d'un système multi-locuteurs. Cette bibliothèque, mise au format DLL (Dynamic Link Library), peut être utilisée par les programmeurs d'applications Windows95, désirant y inclure une interface vocale.
L'application interagit avec deux types « d'interlocuteurs » : l'homme à travers une interface homme-machine, et les autres applications via des messages.
L'apprentissage du vocabulaire peut être fait en cours d'utilisation, évitant ainsi une longue et fastidieuse séance de répétitions au micro pour le sujet.
Le logiciel est capable de reconnaître des mots connectés et il a la capacité d'adaptation à l'usager.

ECHO est composé de 2 modules principaux : le module d'apprentissage (ECHO Lets'TALK ou ECHOTALK) et le module temps réel (ECHO_Lets'GO ou ECHOGO).
ECHOTALK est le logiciel d'apprentissage et de configuration. Avec ECHOTALK on peut définir les caractéristiques de l'utilisateur, les mots et les phrases à reconnaître et les actions associées à chaque phrase reconnue.
ECHOGO tourne en arrière plan et exécute l'action associée à la phrase reconnue. Il peut lancer des applications, envoyer une chaîne ASCII à une autre application, lancer une Macro ou lancer un driver MCI.

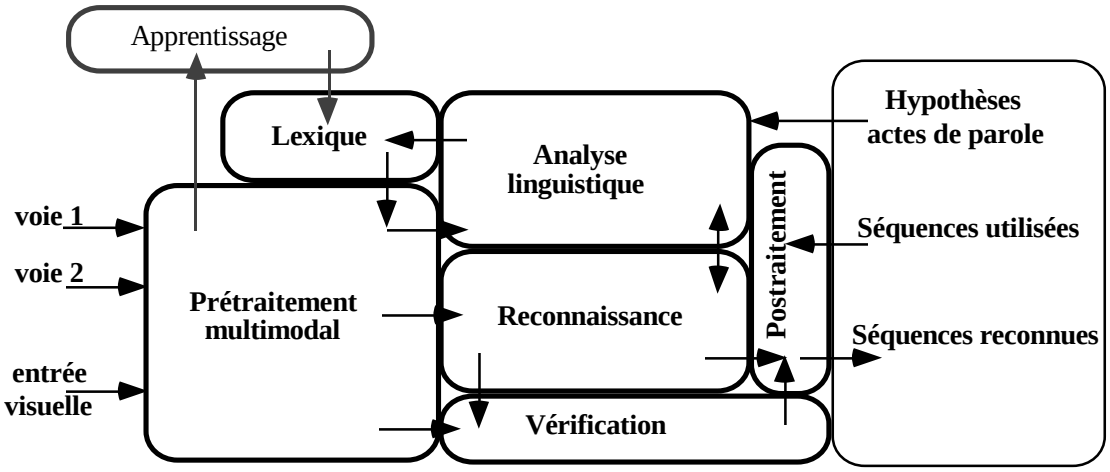


Figure 3 - L'architecture du logiciel ECHO

3-2- Le logiciel FIRAS

FIRAS est un logiciel d'interaction qui accepte en entrée des commandes ayant une syntaxe bien définie et les exécute. Ces commandes concernent des tâches de dessin. L'utilisateur du logiciel a le choix entre l'utilisation de la souris pour dessiner (choisir les objets d'une palette (voir figure 2), les déplacer ou bien les effacer) et l'utilisation du *prompt* des commandes pour dessiner. Il est aussi possible de dessiner en utilisant des commandes multimodales par exemple avec un geste pour designer les lieux (avec le mot *here*) et les objets (avec le mot *this*) par des clicks souris.

Le schéma suivant présente l'architecture du logiciel FIRAS :

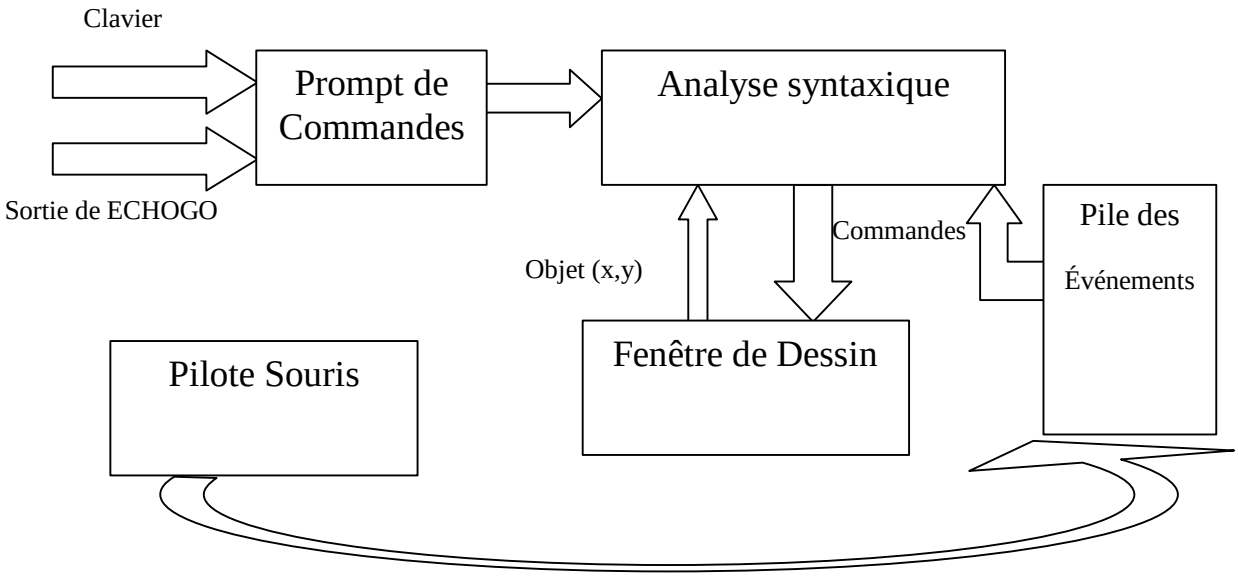


Figure 4 - L'architecture du Logiciel FIRAS

Les modules du logiciel FIRAS ont les fonctions suivantes :

Prompt des Commandes : est une fenêtre qui accepte les commandes tapées sur le clavier ou envoyées depuis le logiciel ECHOGO. Le Logiciel ECHOGO qui tourne en arrière plan a les fonctions de serveur de parole. Les phrases reconnues par ECHOGO sont envoyées sous forme de chaîne de caractères au module *Prompt des Commandes* de FIRAS. Ce module envoie périodiquement les chaînes reçues au module *Analyse syntaxique*.

Analyse syntaxique : est le module « noyau » du logiciel FIRAS. Le module *Analyse syntaxique* est constitué de :
1- Un analyseur lexical,
2- Un analyseur syntaxique,
3- Un fusionneur d'informations multimodales, et
4- Un constructeur de la commande multimodale.

Le module *Analyse syntaxique* reçoit les commandes sortant du *Prompt de Commandes* sous forme de chaîne de caractères et fait l'analyse lexicale, l'analyse syntaxique et la fusion d'informations reçues avant de donner la commande sémantiquement correcte à la fenêtre de dessin. Les informations fusionnées dans ce module pour construire la commande finale proviennent de la pile d'événements et de la fenêtre de dessin ainsi que de l'analyseur syntaxique qui fait partie du module *Analyse syntaxique*. Le fusionneur de FIRAS fait l'association entre les entrées provenant du *Prompt de Commandes* et les objets déjà dessinés sur la

fenêtre de dessin (données provenant de la *Fenêtre de Dessin*) et les événements (clicks souris – données provenant de la *Pile des événements*) afin de construire la sémantique de la commande multimodale.

Pile des événements : est une pile dans laquelle le pilote de la souris insère des données concernant les clicks souris effectués sur la fenêtre de dessin. Ces données spécifient le lieu et le temps de chaque click souris.

Fenêtre de dessin : est un module spécialisé qui accepte des commandes sous forme codée et les exécute. Ce module est un logiciel de dessin qui a une palette d'outil classique dans laquelle on peut choisir de dessiner, de déplacer ou d'effacer un objet (voir figure 2). Les fonctions de ce module sont déclenchées soit par des messages provenant d'un autre processus (*Analyse syntaxique* dans notre cas) et qui contiennent des commandes sous forme codée ou bien par des messages du système d'exploitation envoyés directement en utilisant la souris pour dessiner sur la *Fenêtre de dessin.*

Les modules du logiciel FIRAS tournent sous Windows95 et communiquent entre eux par des messages du système.

La syntaxe 'multimodale' acceptée par le logiciel FIRAS est la suivante :

Action = [draw] [move] [delete]
Objet = [linev] [lineh] [trianglebig] [trianglesmall] [circle] [rectangle] [*déictiques* = this + ***Mouse Click***] [*anaphoriques* = it]
Lieu = [PositionRelative(objet)] [Position] [PositionAutoRelative]
PositionRelative(objet) = [over the(objet)] [under the(objet)] [left to the (objet)] [right to the(objet)]
Position = [here + ***Mouse Click***]
PositionAutoRelative = [more up] [more down] [more left] [more right]

Le logiciel FIRAS a été entièrement développé dans le cadre de cette recherche et pour les buts spécifiques des expériences.

3-3- Propriétés du système

L'environnement de dessin composé du logiciel ECHO et du logiciel

FIRAS supporte l'utilisation de la multimodalité dans quatre contextes : exclusif, concurrent, alterné et synergique.

- *Le contexte exclusif* est supporté autant qu'on peut utiliser un des modes valable pour faire des tâches différentes. On peut dessiner en utilisant la souris seulement, ou le clavier pour entrer des commandes sur le prompt des commandes ou en utilisant la parole seule en s'adressant à ECHOGO qui passe les commandes au prompt des commandes.
- *Le contexte alterné* est supporté autant qu'on peut écrire une commande « *draw* » (clavier) puis designer un lieu (click souris) puis prononcer le reste de la commande « *circle here* » afin de dessiner un cercle.
- *Le contexte concurrent* est supporté autant qu'on peut travailler avec la souris pour dessiner un objet et en même temps donner des commandes vocales pour dessiner un autre objet.
- *Le contexte synergique* est supporté autant qu'on peut prononcer la commande « *draw circle here* » et designer le lieu par un click souris ou qu'on peut prononcer (ou taper) la commande « *move this here* » et designer un objet et un lieu par des click souris.

4- Déroulement de l'expérience

L'expérience s'est déroulée dans les mêmes conditions que pour l'expérience avec l'assistant, mais cette fois-ci avec l'interface multimodale réelle. Les sujets étaient tous familiers des logiciels de dessin. Les 8 sujets testés avaient passé la même expérience avec l'assistant humain (chapitre 4) et ils pouvaient se familiariser avec le logiciel (et même devenir relativement expert) avant l'expérience (1 semaine d'entraînement). Les expériences étaient filmées et analysées manuellement (en utilisant une camera vidéo pour filmer toutes les expériences et les analyses étaient faites ultérieurement sur un écran de télévision) : on a relevé pour chaque action les propriétés CARE des sujets. Les actions retenues par l'analyse sont données dans l'arbre (fig. 5), c'est-à-dire essentiellement les actions à résultat immédiat (les mêmes mesures que pour l'expérience précédente).

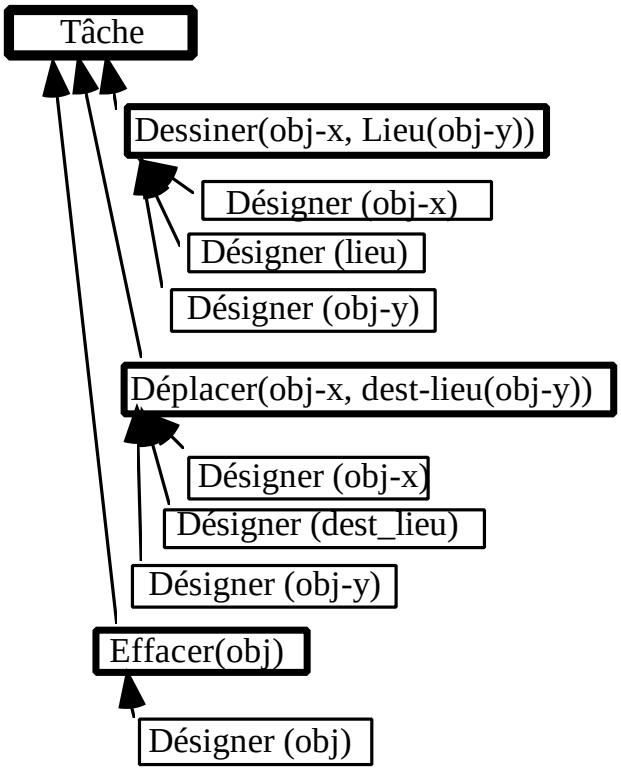


figure 5 : les tâches élémentaires dans l'expérience

5- Résultats tirés de l'expérience

Nous reprenons ci-dessous le même plan de présentation des résultats que dans le chapitre 4 afin de faciliter la lecture comparative des résultats. Il y aura également quelques rappels des définitions. On notera (pdd) les actions multimodales souris+parole.

5-1- L'équivalence des modes

L'équivalence des modes désigne le fait d'obtenir un même résultat en utilisant différentes modalités [17]. On peut distinguer deux notions : l'équivalence de résultat et l'équivalence fonctionnelle.

Equivalence de résultat : les modes proposés dans notre interface (pdd) (p) et (s) sont équivalents vis-à-vis du résultat final (obtenir une figure). En effet chaque mode est suffisant à lui seul pour effectuer la tâche de dessin à accomplir.
Equivalence fonctionnelle : il n'y a pas d'équivalence fonctionnelle entre les modes dans notre interface.

Nous avons relevé le nombre de fois où l'un et/ou l'autre mode ont été utilisés pour l'ensemble des dessins et pour tous les utilisateurs *(par action et non par personne)*. Les résultats sont les suivants :

(pdd)	(p)	(s)
68.5%	24.3%	7.2%

tableau 1 : usage des modes

On constate, à partir de ce tableau (tableau 1), qu'il y a un usage prédominant du mode pdd. L'usage du mode (s) étant occasionnel était caractérisé par l'existence d'une erreur de reconnaissance ou d'une difficulté de trouver la syntaxe convenable facilement (70.8% des cas d'utilisation du mode souris). Par contre, l'usage de la parole (mode p) était favorisé par l'absence de la nécessité de précision (effacer tous les objets sans tenir compte de l'ordre).

Ces résultats confirment ceux de l'expérience avec assistant humain : le mode (pdd) multimodal, qui semblait le plus "naturel" emporte effectivement l'adhésion des sujets qui l'utilisent dans la majorité des cas. On note également une légère augmentation de l'usage du mode parole seul.

5-2- L'assignation des modes

Pour estimer le degré d'adéquation d'un mode à un type d'action (assignation) nous avons calculé pour chaque action et selon son degré de réalisation, le pourcentage relatif des modes utilisés. Avec plus de détails par types d'action nous avons :

"dessiner"		
(s)	(p)	(pdd)
4.75%	16.28%	79.07%
"déplacer"		
(s)	(p)	(pdd)
44.7%	17.3%	38%
"effacer"		
(s)	(p)	(pdd)
0	34%	66%

tableau 2 : usage des modes selon le type d'action pour les cas de réussite de l'action

On peut constater que l'usage du mode (s) est devenu plus important vis-à-vis de l'expérience précédente pour l'action « déplacer » ; ceci est certainement dû à la facilité et à l'efficacité de la souris pour ce genre de tâche (il faut désigner un objet précis et un lieu). Pour les autres actions il y a plutôt une diminution du mode (s). On remarque donc sur cet exemple une nette spécialisation vis-à-vis de la tâche, sans doute sous le double effet, d'une part d'une meilleure adéquation de certains modes aux actions, et d'autre part, des erreurs de reconnaissance de la parole. Il est à

noter toutefois que ces erreurs ne pénalisent pas gravement l'usage de la parole qui reste attractive pour l'utilisateur.

Les cas d'échec de l'action (voir les valeurs dans le tableau ci-dessous) sont peu nombreux par rapport au nombre total d'actions tentées, car les sujets sont devenus familiers avec les défauts du logiciel et ont réussi à s'y adapter en évitant les situations d'impasse. Il est évident que les échecs les plus nombreux viennent du logiciel de reconnaissance de la parole.

"dessiner"		
(s)	(p)	(pdd)
0	10	2
"déplacer"		
(s)	(p)	(pdd)
0	14	3
"effacer"		
(s)	(p)	(pdd)
0	2	1

tableau 3 : usage des modes selon le type d'action dans les cas d'échec de l'action

5-3- La complémentarité et la redondance

La redondance n'est pas prévue dans le logiciel qui n'accepte qu'une syntaxe limitée. Les sujets ayant été longuement entraînés et connaissant donc les limites du système, ils n'ont prononcé que très rarement des commandes redondantes (3.8% de tentatives et le système n'a pas réagi dans ces cas). Pour les actions multimodales (pdd) l'usage de la complémentarité a donc été de 96.2%.

6- Observations générales

Réparation des erreurs

Le système a fait des erreurs dans l'exécution des commandes du fait d'une mauvaise reconnaissance de la commande. Dans tous les cas le sujet a tenté de les réparer, il n'y a pas eu d'abandon.

Réparation monomodale et multimodale : dans 78% des cas d'erreurs l'usager a réitéré la commande et dans 82% des cas le sujet est passé à un autre mode après le deuxième ou troisième échec.
Le mode pdd était renforcé dans l'expérience précédente, il est ici substitué par le mode (s) car la plupart des erreurs a pour origine la

reconnaissance de la parole (le mode initial est le mode (p) ou le mode (pdd) dans 98% des cas).

mode correction	pourcentage
s	53%
p	13%
pdd	34%

tableau 4 : fréquence d'utilisation des modes dans la phase de réparation

7-Conclusion et prolongements

Les résultats de l'expérience donnent une première idée sur le comportement multimodal de l'utilisateur dans cette situation de communication homme-machine réelle. Par rapport aux résultats de l'expérience précédente où l'usager était en communication humaine on peut constater que l'usage de la multimodalité semble globalement plus important, malgré les insuffisances des performances du système de reconnaissance.

La correction des erreurs repose beaucoup sur le mode (s) classique et plus sûr.

En ce qui concerne les propriétés CARE, on note que (a) il y a une tendance pour le sujet à une assignation naturelle des modes selon le type de commande et que (b) bien que la redondance n'ait pas été implémentée dans le logiciel, il ne semble pas que cette absence ait été une gêne pour l'usager car la tentation de l'utiliser a été faible après la phase d'apprentissage.

Nous pouvons également risquer une interprétation plus générale malgré les limitations de notre expérience : l'usage des modes est fondé sur la fiabilité et le risque d'échec, ce qui se traduit par le fait qu'après une première tentative « naturelle » sur un mode, l'utilisateur passe à un autre mode plus fiable pour rattraper un échec.

Une conclusion importante retrospectivement est que la situation simulée avec l'assistant (à faible coût de développement et de mise en place) donne déjà des résultats significatifs sur l'usage de la future interface. C'est encourageant pour l'introduction d'une telle méthode simplificatrice,

comme procédure prédictive pour l'accompagnement de la conception de l'interface.

Chapitre 6

Conclusions sur l'étude expérimentale sur l'usage de la multimodalité

En termes généraux, la multimodalité apparaît plus comme une multimodalité actionnelle que comme une multimodalité informationnelle. Nous pensons donc que d'autres critères que les critères CARE doivent être pris en compte pour juger de la multimodalité car ils sont trop liés à la notion d'information (de manière similaire à la théorie de Bernsen [3]) et pas assez à celle d'action ou d'interaction (par exemple pour juger de la rémanence d'un processus actionnel ou de l'efficacité d'une action). Pour cela on peut référer aux travaux de Zanello, Caelen et Bisseret sur les critères T-CCARE [24].

Notre étude est restée basée sur les propriétés CARE qui sont présentées dans le chapitre 2. Les conditions de ces deux expériences, les logiciels utilisés et les outils de mesure nous ont permis de faire l'étude au niveau de l'acte et non au niveau de la tâche

Les raisons d'utilisation de tel ou tel mode sont à préciser par de nouvelles expériences plus ciblées et sur des situations plus variées. Le travail doit donc être poursuivi pour mieux comprendre les comportements des usagers face à une interface multimodale ; mais déjà, à ce stade de notre investigation nous avons pu juger positivement de l'intérêt des propriétés CARE pour situer l'usager par rapport à un contexte d'interaction multimodale, ainsi que de l'intérêt de méthodes de simulation d'usage.

Section 2
Annexe 1

Bibliographie de la section 2

[1] Bell D., Johson P., General models of multimedia interaction. ERCIM Workshop Report 94-W003, Inria, Nancy, p. 25-39, 1994.

[2] Bellik Y., Interfaces multimodales : concepts, modèles et architectures. Thèse de l'Université d'Orsay, LIMSI, 1995.

[3] Bernsen N.O., Modality theory : supporting multimodal interface design.ERCIM Workshop Report 94-W003, Inria, Nancy, p. 13-23, 1994.

[4] Bretan Y., Karlgren J., Synergy effects in natural language-based multimodal interfaces. ERCIM Workshop Report 94-W003, Inria, Nancy, p. 43-58, 1994.

[5] Bisson P., Nogier J-F., Interaction homme-machine multimodale : le système MELODIA dans "ERGO-IA92" Biarritz, p. 69-90, 1992.

[6] Bourguet M.L., Conception et réalisation d'une interface de dialogue personne-machine multimodale, thèse de doctorat, ICP/INPG, Grenoble 1992.

[7] Caelen J., Coutaz J., Interaction homme-machine multimodale: quelques problèmes. IHM'91, Troisièmes journées sur l'ingénierie des interfaces homme-machine. Dourdan 11-13 December 1991.

[8] Caelen J., Compte rendu du workshop IHMM organisé par le GDR-PRC "Communication Homme-Machine", Dourdan 13-14 April, p. 213-228, 1992.

[9] Caelen J., Multimodal Human-Computer Interface, First Al-SHAM International Conference on Information Technology, Damascus - Syria , May 1994.

[10] Carbonell N., Valot Cl., Mignot Ch., Dauchy P., Etude empirique : usage du geste et de la parole en situation de communication homme-machine. Actes du congrès ERGO'IA, p. 128-139, 1994.

[11] Catinis L., Caelen J., Multimodal Man-Machine Interaction in Administration and Public Services, First Al-SHAM International Conference on Information Technology, Damascus - Syria , May 1994.

[12] Chapelier L., Fay-Varnier Ch., Roussalany A., Saint-Dizier V., Recueil et analyse d'un corpus d'interactions multimodales homme-machine. Actes du congrès ERGO'IA, p. 96-107, 1994.

[13] Coutaz J. & Gourdol A., Communication homme-machine multimodale: perspectives pour la recherche. Deuxièmes Journées Nationales GRECO PRC CHM, Toulouse, p. 17-28, January 1991.

[14] Coutaz J. et Nigay L., Les propriétés "CARE" dans les interfaces multimodales, IHM'94 Sixièmes journées sur l'ingénierie des interface Homme-Machine. Lille, p. 7-14, 8-9 décembre 1994.

[15] Faure C., Julia L. , Interaction homme-machine par la parole et le geste pour l'édition de documents : TAPAGE. Actes interface des mondes réels et virtuels. Montpellier, p. 71-180, 1994.

[16] Faure C. ,Arnold M. , L'interaction homme-machine du point de vue des principes d'économie. Actes de cinquièmes journées sur l'ingénierie des interface Homme-Machine IHM'93, Lyon, p. 3-8, 19-20 Octobre 1994.

[17] IHM'93, Synthèse de l'atelier "Interfaces Multimodales", sous-groupe : "Formes de multimodalité en situation d'interaction utilisateur-machine".

[18] Martin J-C , Béroule D., Types et buts de coopération entre modalités, IHM'93, 5ème journées sur l'ingénierie des interfaces homme-machine, Lyon, p. 17-22, 1993.

[19] Martin J-C , Etude fondée sur des Types et buts de coopération entre modalités. Troisièmes journées internationales "L'interface des mondes réels et virtuels". Montpellier 1994.

[20] Mignot Ch., Valot Cl., Carbonell N., An experimental study of future "natural" multimodal human-computer interaction. Proc. InterCHI'93, Amsterdam, p. 67-68, 1993.

[21] Ozkan, N., Caelen J., Designation of graphical objects in human-computer interaction. Proc. of WWDU'92 congresss. Berlin, 1992.

[22] Valot Cl., Interface Multimodale Projet Grenoblois, Rapport d'activité du pôle Interfaces Homme-Machine Mulitmodales, GDR-PRC CHM, 1993.

[23] E. Brison, N. Vigouroux, Interprétation des événements dans l'interaction multimodale. Actes IHM'94, p. 23-28, 1994.

[24] Zanello M.L. ; Caelen J. ; Bisseret A. (1996). Une approche centrée tâche de la multimodalité, IHM'96, p99-105.

Section 3

L'utilisabilité de la multimodalité

Chapitre 7

Etude sur l'utilisabilité de la multimodalité

1. Introduction : l'utilisabilité de la multimodalité
2. Les critères ergonomiques de l'utilisabilité
3. L'étude expérimentale - but et objectif
4. Une selection des critères ergonomiques pour la multimodalité
5. Le système de commande vocale utilisé
6. Les sujets de test
7. Le déroulement de l'expérience
8. Les mesures et les résultats obtenus
9. Observations et commentaires
10. Conclusion et prolongements

1- Introduction : l'utilisabilité de la multimodalité

L'un des but des recherches dans le domaine de la multimodalité est la valorisation des modalités nouvelles dans les interfaces homme-machine. La parole est une des modalités candidates des nouvelles interfaces homme-machine car il existe actuellement sur le marché des micro-ordinateurs, plusieurs logiciels de reconnaissance de la parole (mots isolés et parole continue). Un exemple de ces logiciels est le logiciel « Microsoft Voice » qui tourne sous « Windows95 » et qui accepte des commandes parlées équivalentes à des commandes de l'interface classique « Windows95 ». Etant donné l'existence de tels outils, il est judicieux de comparer l'utilisation des interfaces multimodales avec l'utilisation des interfaces graphiques classiques.

Pour une évaluation en vraie grandeur de ces interfaces, il faudrait sortir de l'environnement de *laboratoire* et aller observer les usagers sur leurs lieux de travail habituels. Comme ces interfaces ne sont pas d'un usage courant en milieu professionnel, nous avons encore mené l'étude en milieu de laboratoire mais avec des sujets experts.

2- Les critères ergonomiques de l'utilisabilité

Plusieurs méthodes d'évaluation des interfaces utilisateur sont disponibles actuellement [3] [4] [5] [6] [7] [9] [10] [11] [12]. On peut les classer en deux grandes catégories [2].

1. La première catégorie considère que l'utilisateur est la source des données de l'évaluation. L'approche objective est l'observation du comportement de l'utilisateur pendant l'interaction réelle, tandis que l'approche subjective est faite en recueillant les avis des utilisateurs par des questionnaires et des entretiens. Ces deux approches peuvent être faites séparément ou conjointement.
2. La deuxième catégorie de méthodes est fondée sur une forme d'inspection par un ergonome. Elle ne nécessite pas obligatoirement, la présence d'utilisateurs.

C'est dans le but de définir une méthode d'inspection que des critères ergonomiques ont été proposés par Bastien et Scapin [1]. L'objectif initial de leurs travaux a été de formaliser et de structurer les critères pour aboutir à des recommandations. Les critères ergonomiques représentent d'abord un moyen de classification des recommandations mais surtout, ils représentent les dimensions ergonomiques majeures selon lesquelles un système interactif peut-être évalué ou spécifié.

La liste des critères ergonomiques de Bastien et Scapin [1] est :

1. *Guidage*
   - *1.1. Incitation*
   - *1.2. Groupement/Distinction entre Items*
     - *1.2.1. Groupement/Distinction par la localisation*
     - *1.2.2. Groupement/Distinction par le format*
   - *1.3. Feed-back Immédiat*
   - *1.4. Lisibilité*
2. *Charge de Travail*
   - *2.1. Brièveté*
     - *2.1.1. Concision*
     - *2.1.2. Actions Minimales*
   - *2.2. Densité informationnelle*
3. *Contrôle Explicite*
   - *3.1. Action Explicites*
   - *3.2. Contrôle Utilisateur*
4. *Adaptabilité*
   - *4.1. Flexibilité*
   - *4.2. Prise en compte de l'expérience de l'utilisateur*
5. *Gestion des erreurs*
   - *5.1. Protection contre les erreurs*
   - *5.2. Qualité des messages d'erreur*
   - *5.3. Correction des erreurs*
6. *Homogénéité/Cohérence (consistance)*
7. *Signifiance des Codes et Dénominations*
8. *Compatibilité*

3- L'étude expérimentale - but et objectif

Le but principal de notre expérience est de comparer l'utilisabilité de la multimodalité dans une application d'édition de textes MSWord « Microsoft Word for Windows » avec et sans multimodalité. MSWord a toutes les facilités des interfaces graphiques « *WYSIWYG*». MSWord supporte des possibilités de manipulation par la souris et des « accélérateurs clavier », ce qui rend le mode gestuel très favorable surtout chez les utilisateurs experts de cet éditeur.

Dans cette expérience nous avons donné des capacités multimodales à MSWord en activant le logiciel « Microsoft Voice » et nous avons demandé aux usagers sujets du test, d'effectuer une tâche dont ils ont l'habitude ; il s'agissait de modifier ou d'écrire une lettre selon un modèle établi à l'avance (voir annexe). Leur comportement a été observé et analysé.

Les mesures faites concernaient la fréquence d'utilisation des commandes vocales, le contexte dans lequel elles étaient utilisées (propriétés CARE), le taux des erreurs de reconnaissance et leur impact sur l'utilisation des modes, et enfin le gain en performance.

Les réactions des sujets et leurs commentaires ont servi - en conjonction avec les statistiques tirées - à donner de nouvelles orientations et de nouvelles idées pour de futures expériences.

Nous avons utilisé une partie des critères ergonomiques de Bastien et Scapin (appelés ci-après critères BS).

4- Une sélection des critères ergonomiques pour la multimodalité

Les critères ergonomiques BS sont applicables sur toutes les interfaces utilisateur connues. L'introduction de la multimodalité a des effets sur les interfaces, et ces effets portent sur certains critères d'une façon prédominante qu'il faut examiner. Dans cette expérience, et en considérant les circonstances sous lesquelles elle s'est déroulée, nous avons constaté que les critères suivants peuvent être considérés : :

- Feedback immédiat,
- Brièveté (Concision et Actions Minimales),
- Adaptabilité (Flexibilité).

Les autres critères ne sont pas intéressants dans notre cas, car ils concernent les modalités en sortie (guidage : incitation, groupement/distinction entre items, lisibilité, charge de travail : densité informationnelle) que l'introduction du logiciel « Microsoft Voice » ne modifie pas.

Le feedback immédiat d'une entrée parlée se fait par un message qui informe l'utilisateur sur la commande reconnue, puis cette commande s'exécute s'il n'y a pas d'erreur. On peut donc dire qu'il y a une sorte de double feedback, mais on ne peut pas dire que la qualité du feedback immédiat soit améliorée *a priori*.

Les commandes parlées acceptées par le système sont de longueur limitée (un ou deux mots maximum) et elles sont équivalentes dans la plus part des cas à une série d'actions gestuelles, ce qui rend *a priori* plus riche le critère de brièveté et ses sous critères (concision et actions minimales).

L'adaptabilité et surtout la flexibilité devraient être améliorées grâce au choix rendu possible entre plusieurs modalités et aux différentes façons d'organiser la tâche.

Avec ces remarques générales, nous allons discuter maintenant des résultats obtenus après avoir détaillé l'expérience et les moyens mis en œuvre.

5- Le système de commande vocale utilisé :

Le système utilisé pour la reconnaissance de la parole et l'exécution des commandes vocale est le logiciel « Microsoft Voice » qui tourne sous « Windows95 » en arrière plan. Une fois « Microsoft Voice » lancé, il faut préciser l'identité du locuteur (déjà défini ou nouveau locuteur). Si c'est le cas d'un nouveau locuteur « Microsoft Voice » demande à faire apprendre sa voix. Le processus d'apprentissage est alors guidé par un « Wizard ». Cela consiste à lire une liste de mots et de phrases pré-définies. L'apprentissage peut être court (session de 5 minutes) ou prolongé (session de 15 minutes). Il faut ensuite choisir le mode de travail parmi les trois modes suivants : (stop listening - pause listening - start listening).

1. Stop Listening : « Microsoft Voice » est rendu inactif,
2. Start Listening : « Microsoft Voice » est rendu actif et les commandes vocales prononcées par le locuteur seront reconnues (si elles font partie de la liste de commandes définies),
3. Pause Listening : « Microsoft Voice » est en attente d'une commande d'activation ou d'un déplacement de souris « Hot Corner ».

Les commandes reconnues par « Microsoft Voice » sont exécutées dans l'application active de « Windows95 », ou elles seront adressées directement au « Program Manager » du système. On peut alors choisir MSWord comme application receptrice, qui devient une application multimodale qui accepte en entrée les commandes provenant de la souris, du clavier et du microphone et supporte les modes d'interaction (geste et parole) associés à ces médias, dans les contextes : exclusif, concurrent, alterné et synergique.

- *Le contexte exclusif* est supporté dans la mesure où on peut utiliser un des modes pour faire des tâches différentes.
- *Le contexte alterné* est supporté dans la mesure où on peut par exemple, écrire un mot (clavier) puis marquer le mot (souris) puis prononcer la commande « Cut Selection » afin de couper ce mot.
- *Le contexte concurrent* n'est pas supporté en principe car on ne peut pas travailler dans une fenêtre et donner des commandes dans une autre parce que le logiciel « Microsoft Voice » s'adresse uniquement à l'application active.
- *Le contexte synergique* est supporté virtuellement dans la mesure où on peut marquer un texte en disant «Copy Selection» et le système fera la copie du texte marqué. En réalité on profite ici du délai dû à la reconnaissance de la parole pour avoir un pseudo-parallélisme (figure 7-1).

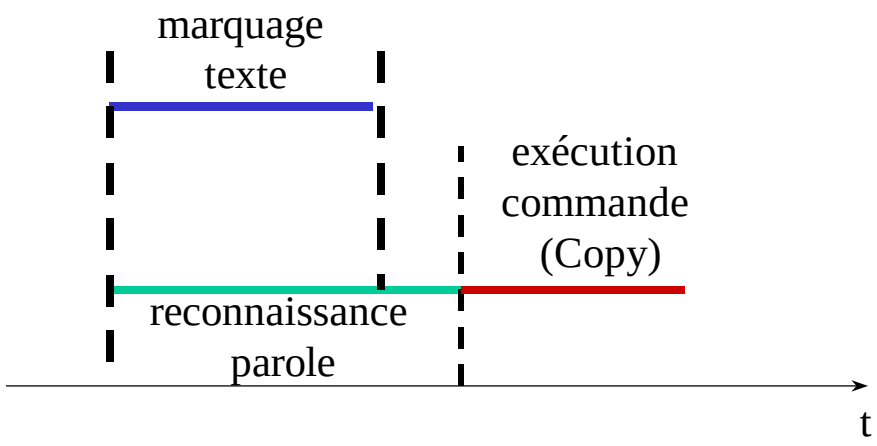


figure 7-1

« Microsoft Voice » dans cette expérience accepte à l'entrée les 40 commandes suivantes (commandes pré-définies) :

Cascade All Windows
Close Window
Copy Selection
Cut Selection
Delete Selection
Escape
List Voice Commands
Maximize Window
Minimize All Windows
Minimize Window
Move Down
Move Left
Move Right
Move Up
Next Field
Next Window
Page Down
Page Up
Paste Selection
Pause Listening
Press Apply
Press Cancel
Press Done
Press Enter
Press No
Press Space
Press Tab
Press Yes
Previous Field
Previous Window
Restore Window
Select Line
Select Word
Show Help
Show Help On This
Start Listening
Stop Listening
Tile All Windows
Undo

Le logiciel « Microsoft Voice » supporte le possibilité d'ajouter de nouvelles commandes. Il supporte aussi la possibilité de programmer des macro-commandes.

6- Les sujets de test

Dans cette expérience 24 sujets de test ont passé l'expérience. Les sujets sont des ingénieurs, des étudiants en sciences, des administratives et des secrétaires.

Les sujets de test sont des experts qui utilisent MSWord d'une façon très régulière. L'âge des sujets va de 22 ans à 30 ans et leur expérience en informatique de bureautique oscille entre 3 ans et 8 ans.

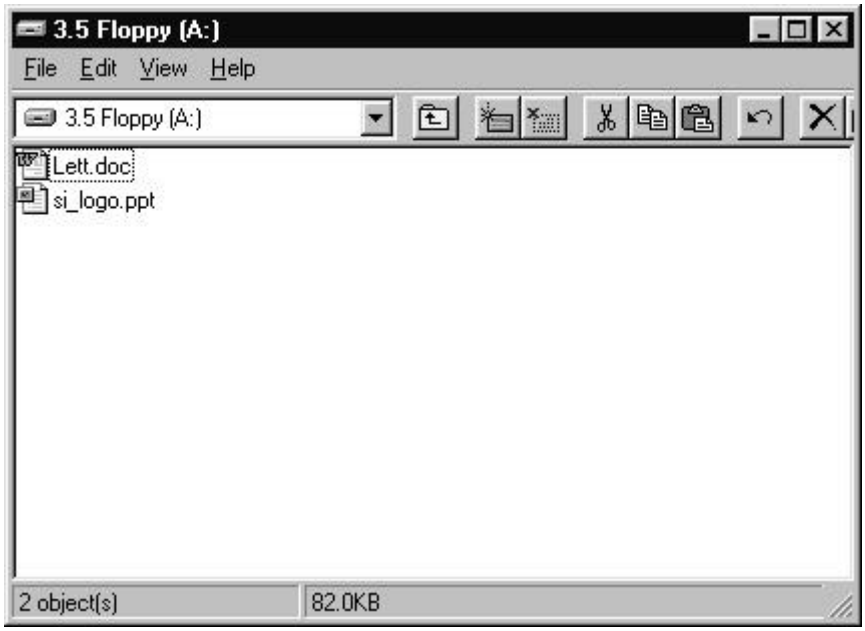
Le choix de sujets de test avait comme objectif d'avoir une variété de sujets et de domaine d'activité pour pouvoir généraliser autant que faire se peut pour tout type d'usager et de domaine d'activité. Par contre les facteurs d'âge, d'expérience et de capacité d'apprentissage étaient pris en compte pour que les résultats aient un sens pour la multimodalité.

7- Le déroulement de l'expérience

L'expérience se déroulait dans un bureau sur un ordinateur personnel PC sur lequel tournait « Windows95 », «Microsoft Word », « Microsoft Power Point » et « Microsoft Voice ». L'expérience était filmée (image et son) par une camera vidéo fixée au-dessus de l'écran de l'ordinateur et les mesures étaient prises à partir du film.

L'usager a le texte d'une lettre sur une disquette et un dessin (logo) sous format (.ppt) pour « Microsoft Power Point ». Il dispose de la correction de la lettre faite à la main sur une feuille imprimée. Sa tâche est de corriger la version électronique de la lettre au vu des corrections manuelles ou d'écrire sur ordinateur une nouvelle version de la lettre à envoyer.

Au départ la situation se présente comme suit :



- L'usager doit :

⇒Copier la lettre sur un nouveau répertoire sur le disque dur
⇒ Ouvrir la lettre par MSWord
⇒ Corriger la lettre
⇒ Insérer le logo (après avoir ouvert le fichier qui contient le logo enregistré sous le format « Microsoft Power Point »)
⇒ Imprimer la lettre
⇒ Recopier la version corrigée de la lettre sur la disquette

- L'usager doit au préalable faire apprendre « Microsoft Voice » à reconnaître sa voix (il a tout son temps pour s'entraîner et apprendre à utiliser les commandes vocales).
- Et enfin, l'usager doit effectuer l'expérience deux fois : une fois sans l'utilisation de la parole et une deuxième fois avec la parole.

L'usager a la liste de commandes vocales et il n'est pas autorisé à créer des macros (cette expérience ayant comme but d'observer l'utilisabilité de la parole dans des actions élémentaires, l'utilisateur aurait pu être tenté par l'addition de nouvelles commandes pour accélérer son travail). Par cette limitation, le critère ergonomique *brièveté* et ses sous critères (*concision* et *actions minimales*) étaient limités pour qu'ils ne soient pas dominants d'une façon qui influence les observations des autres critères.

8- Les mesures et les résultats obtenus

Les mesures tirées du film sont les suivantes :

a) Le nombre de fois où l'usager a utilisé une commande vocale.
b) Le type d'utilisation des modalités (synergique ou alterné).
c) Le nombre de fois où le système n'a pas reconnu la commande vocale et la réaction de l'usager dans ce cas.
d) Les commandes que l'usager a entrées sans savoir que le système n'est pas capable de les reconnaître.
e) Le temps global de la tâche.
f) La différence de temps nécessaire pour accomplir la tâche avec et sans utilisation de la multimodalité.

Les résultats sont les suivants :

I- Les commandes utilisées et le nombre des erreurs :

a- Le nombre d'utilisateurs qui n'ont pas utilisé la parole est de 4/24 (16.67 %). Ces utilisateurs ont avancé des raisons différentes comme « ayant l'habitude d'utiliser le clavier et la souris et étant bien convaincus que l'introduction des

modalités nouvelles ne leur changera pas cette habitude », où ils ne sont pas « convaincus que la parole pourrait être plus efficace ou plus rapide que l'utilisation classique du clavier et de la souris dans ce genre d'application ». Enfin il y avait des opinions plus agressives qui disaient : « on n'est pas fous pour parler à l'ordinateur ».

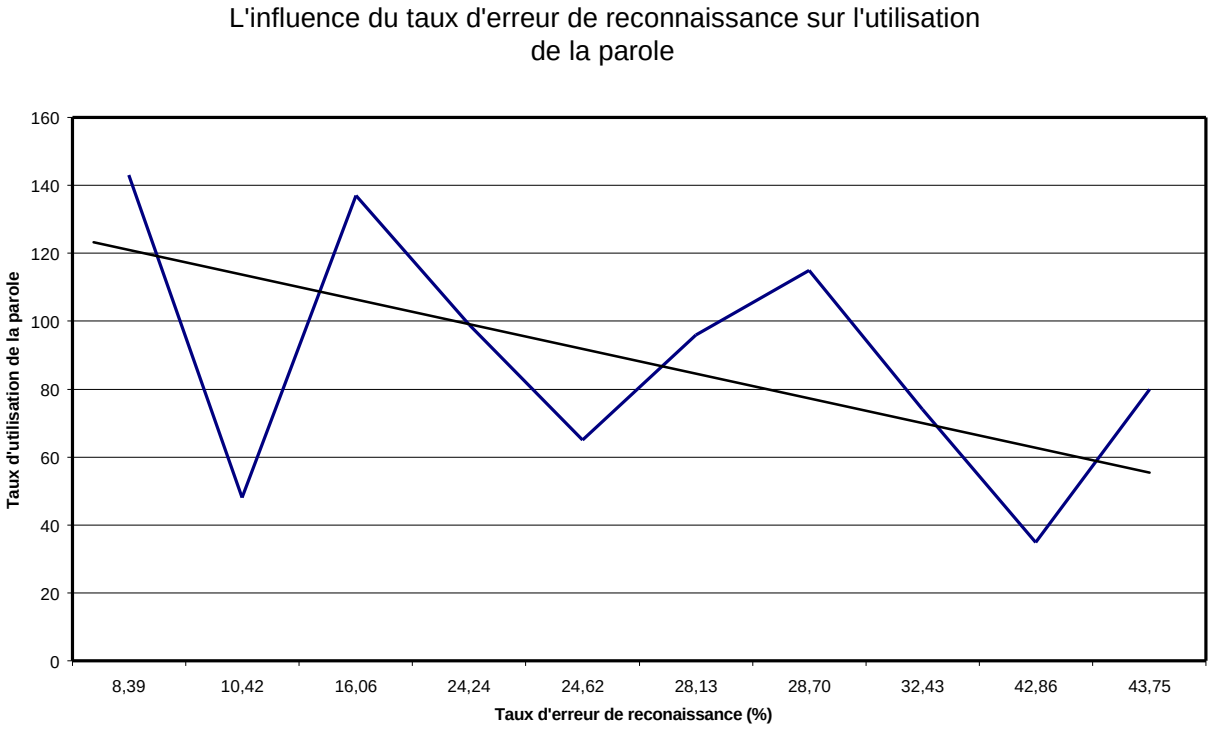
b- Le nombre des commandes vocales utilisées dans cette expérience par tous les sujets est de 1784 commandes. Le tableau suivant montre le taux d'utilisation de chaque commande dans le cadre de l'expérience.

Commande	Nombre d'utilisation	Taux d'utilisation vis-à-vis du total des commandes utilisées
Close Window	44	2.47%
Copy Selection	50	2.80%
Cut Selection	134	7.51%
Delete Selection	36	2.02%
Minimize Window	12	0.67%
Move Down	346	19.39%
Move Left	104	5.83%
Move Right	130	7.29%
Move Up	112	6.28%
Next Field	40	2.24%
Next Window	6	0.34%
Page Down	81	4.54%
Page Up	127	7.12%
Paste Selection	104	5.83%
Press Cancel	8	0.45%
Press Enter	79	4.43%
Press No	31	1.74%
Press Space	17	0.95%
Press Tab	8	0.45%
Press Yes	61	3.42%
Previous Field	2	0.11%
Previous Window	2	0.11%
Select Line	67	3.76%
Select Word	113	6.33%
Undo	70	3.92%
Total	***1784***	***100%***

Il faut bien constater que l'utilisation des commandes est très dépendante de la nature des tâches à accomplir. Il y a des commandes qui n'ont jamais été utilisées dans cette expérience comme les commandes «*Cascade all windows*» et «*Minimize all windows*». Par contre il y avait des commandes peu utilisées à cause de l'existence des raccourcis ou des équivalents gestuels plus faciles ou plus efficaces comme les commandes "*Press Space*, *Escape*", etc. Les commandes les plus utilisées ont été les déplacement de curseur, la sélection et le couper-coller.

c- Le nombre de fois où le système de reconnaissance n'a pas reconnu la commande vocale est de 426 ce qui donne un pourcentage d'erreur en reconnaissance de 23.88 %. Les erreurs sont dûes au bruit de l'environnement, à la mauvaise prononciation des commandes vocales (la langue maternelle des sujets n'est pas l'anglais qui est utilisé dans «Microsoft Voice»), à la position inconfortable du microphone et aux performances du matériel et du logiciel en général.

d- On peut constater que le taux d'utilisation de la parole au cours de l'exécution de la tâche est influencé par le taux d'erreur de reconnaissance : il est évident que ce taux décroit puisque l'utilisateur peut se décourager devant des erreurs de reconnaissance trop fréquentes. La courbe ci-dessous montre la relation entre le taux d'utilisation de la parole et le taux des erreurs chez les sujets.



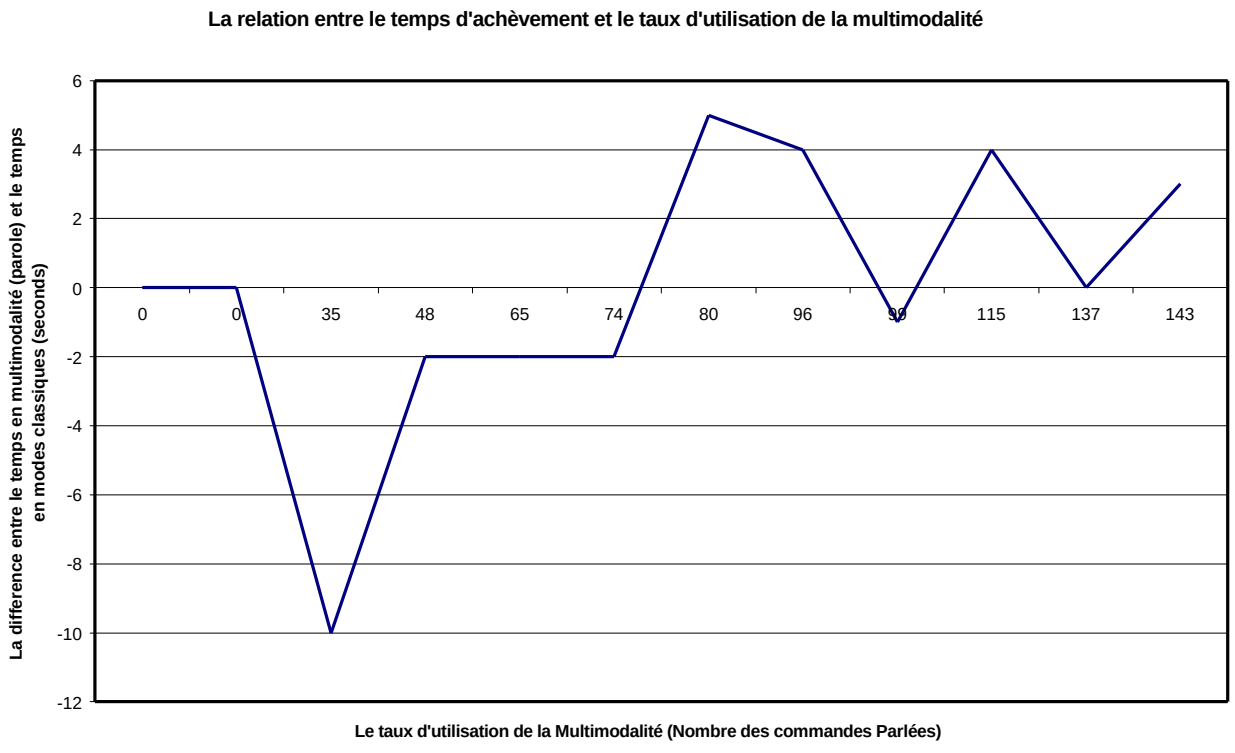
e- Les sujets de test ont utilisé des commandes qui n'existent pas dans la liste des commandes reconnues par le système. On peut en déduire que l'introduction de ces nouvelles commandes pourrait améliorer l'utilisabilité du système multimodal. Les commandes que les sujets ont essayées sont : (*Home, End, Open File, Insert Table, etc.*). Les sujets ont prononcé très souvent les commandes composées de deux mots en abrégeant au premier mot tellement cela leur paraissait évident, comme « *Cut* » à la place de « *Cut Selection* » ou « *Delete* » à la place de « *Delete Selection* » ou « *Down* » à la

place de « *Move Down* ». Ceci donne un indice sur l'importance de la brièveté et de la compatibilité (l'usager est habitué à selectioner « *Edit-Cut* » du menu de Windows95 et non « *Edit-Cut Selection* »).

II- La durée de réalisation de la tâche par l'usager :

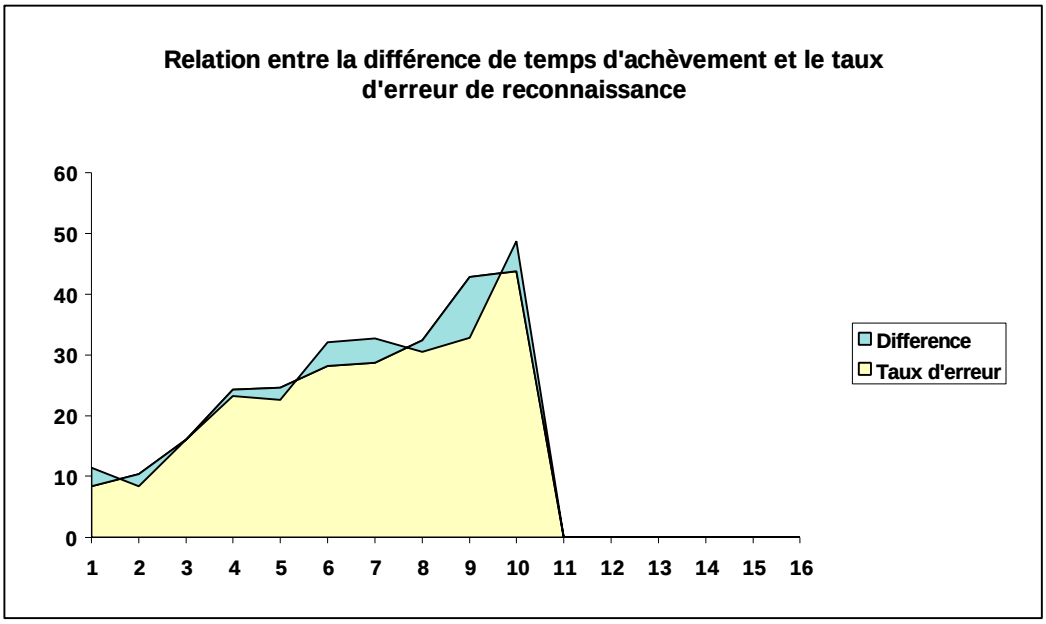
a- Il y a une différence de durée de réalisation de la tâche entre les deux essais (avec la parole et sans la parole). Cette différence est dans quelques cas positive (le temps avec la parole est plus important) mais dans la plupart des cas elle est négative (la durée avec la parole est moins grande). La différence relative positive maximale est de 20%. La différence négative maximale (ou gain) est de 30%.
Les sujets qui ont amélioré leur perfomance sont les usagers qui se sont le plus entraînés à l'utilisation de la multimodalité. Cependant il existe des cas où l'usager était bien entraîné à la multimodalité mais où il n'obtient pas de bonnes performances : on peut avancer deux raisons pour expliquer cela (a) soit qu'il garde une habileté supérieure pour les modalités classiques, (b) soit qu'il a des difficultés de coordination parole/geste.

b- La relation entre le gain et l'utilisation des commandes parlées est indiquée dans la courbe suivante :



On peut remarquer qu'il n'y a pas une relation claire et que le gain est beaucoup plus dépendant de l'usager et de ses caractéristiques propres.

c- La relation entre la durée et le taux d'erreur de reconnaissance des commandes parlées est indiquée sur la courbe suivante :



Il est clair que les erreurs de reconnaissance influencent directement la performance globale de l'utilisation de la multimodalité. Plus les erreurs de reconnaissance sont nombreuses, plus de temps pour accomplir la même tâche est long.

III- L'utilisation synergique et l'utilisation alternée :

L'utilisation de base du système multimodal est « alternée » de part les caractéristiques du système utilisé. Mais dans certains cas, quelques usagers étaient capables d'exploiter intelligemment les caractéristiques pseudo-synergiques du système multimodal. Le pourcentage de cette utilisation synergique est de 3.5 % de la totalité des utilisations de la parole. Une bonne partie de cette utilisation (plus de 60%) était spontanée (les sujets n'ont pas étudié les caractéristiques du système pour prendre conscience de cette possibilité).

IV- La correction des erreurs et la réaction des usagers :

Dans les cas d'erreur de reconnaissance, les usagers avaient plusieurs sortes de réactions :

1. L'usager répète la commande vocale.

2. L'usager passe à un mode diffèrent. Le passage de la parole à cet autre mode était fait après :
    - un essai non réussi,
    - deux essais non réussis, ou
    - après plus de deux essais non réussis.

Le tableau suivant montre les stratégies des usagers en cas d'erreur de reconnaissance :

Taux de changement de mode en cas d'erreur	65%
Taux de changement de mode après la première erreur (un essai non réussi)	35%
Taux de changement de mode après la deuxième erreur (deux essais non réussis)	17%
Taux de changement de mode après plus de deux essais non réussis	13%
Taux de continuation dans le même mode jusqu'à la réussite	35%

On peut constater que l'usager passe à un mode diffèrent dans la plupart des cas. Ceci indique que la performance et la fiabilité d'une modalité influence fortement son utilisation.

9- Observations et commentaires

Après la fin de chaque expérience les usagers étaient questionnés pour connaître leurs impressions et leurs commentaires. On a obtenu les commentaires suivants :

- effet de confidentialité : quelques usagers ne se trouvent pas à l'aise en utilisant la parole et ils préfèrent utiliser le clavier ou la souris bien que la parole facilite les choses - à leur avis - dans beaucoup de cas.
- recherche d'efficacité : si la performance du système multimodale était meilleure (reconnaissance plus rapide et plus fiable) l'usager aurait davantage utilisé la multimodalité.

- recherche de commodité : si l'usager avait pu ajouter de nouvelles commandes au système, il aurait eu plus de raisons pour utiliser la multimodalité (cette option est possible dans « *Microsoft Voice* » mais elle était interdite au cours de cette expérience).
- effet de curiosité : quelques usagers étaient attirés par la possibilité d'utiliser des modalités nouvelles.

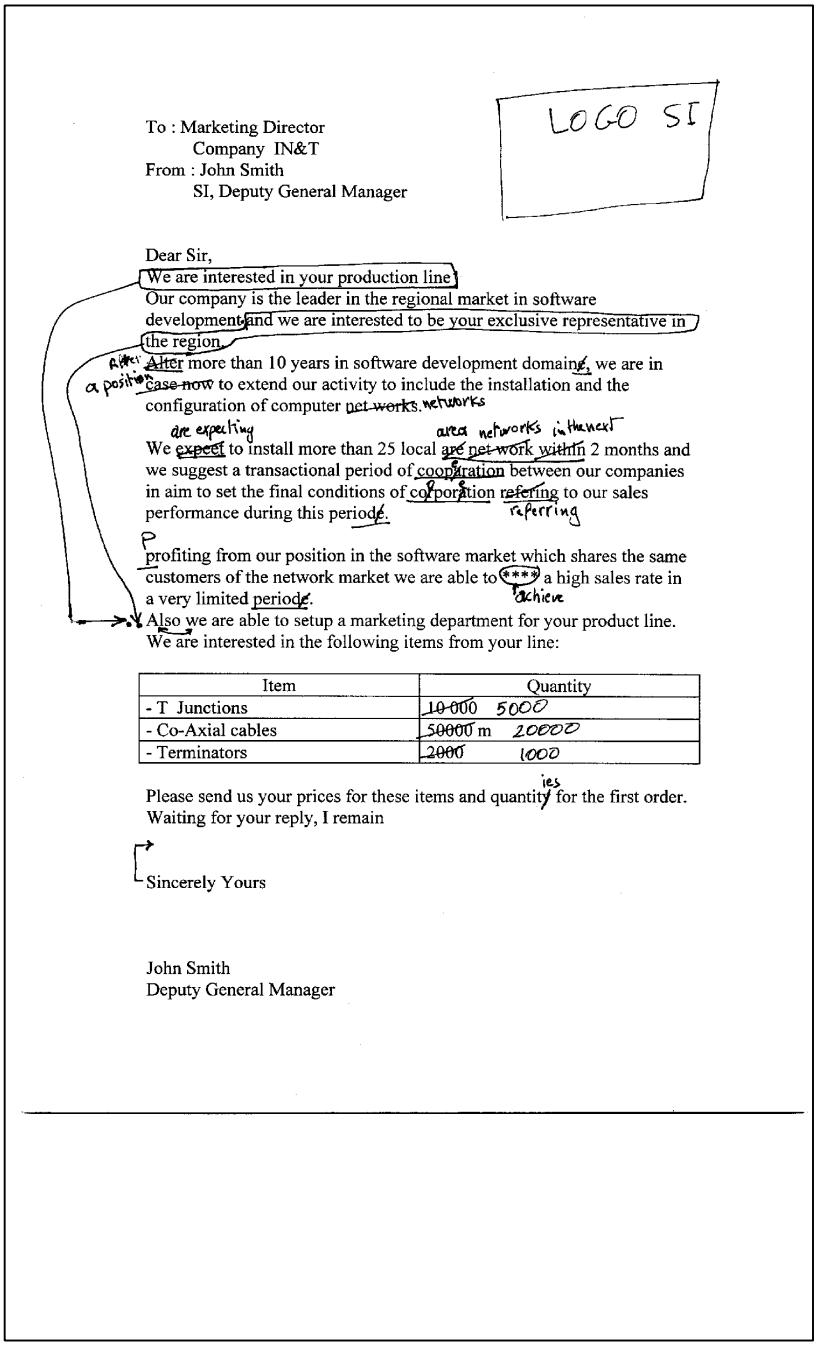
10- Conclusion et prolongement

Les résultats de cette expérience donnent de bons indices sur l'utilisabilité de la multimodalité. Les critères « brièveté » et « adaptabilité » sont maintenant mieux satisfaits et les effets de cette amélioration sont perceptibles à l'usager. Par contre, la fiabilité de l'interface homme-machine est mise en cause parce que le taux d'erreurs reste encore élevé : c'est la raison principale qui a limité l'utilisation efficace de la multimodalité.

Enfin, la durée d'entraînement pour l'apprentissage de la multimodalité était relativement courte. L'utilisabilité étant influencée par les habitudes d'utilisation, l'usage de la multimodalité était donc pénalisé. Un prolongement possible de cette expérience est d'installer « *Microsoft Voice* » sur les ordinateurs de nombreux usagers, de les entraîner, et de leurs demander de l'utiliser pendant une longue période (3-4 mois) et de répéter la même expérience (taper et corriger des lettres) à la fin de cette période.

Section 3
Annexe 1

La lettre utilisée dans l'expérience



To : Marketing Director
Company IN&T
From : John Smith
SI, Deputy General Manager

LOGO SI

Dear Sir,

We are interested in your production line.
Our company is the leader in the regional market in software development and we are interested to be your exclusive representative in the region.

After more than 10 years in software development domaine, we are in a position case now to extend our activity to include the installation and the configuration of computer net works. networks

We expect to install more than 25 local are net work within 2 months and we suggest a transactional period of cooparation between our companies in aim to set the final conditions of corporation refering to our sales performance during this periode.

profiting from our position in the software market which shares the same customers of the network market we are able to **** a high sales rate in a very limited periode.

Also we are able to setup a marketing department for your product line.

We are interested in the following items from your line:

Item	Quantity
- T Junctions	10 000 5000
- Co-Axial cables	50000 m 20000
- Terminators	2000 1000

Please send us your prices for these items and quantity for the first order.
Waiting for your reply, I remain

Sincerely Yours

John Smith
Deputy General Manager

La lettre de la deuxième étape (la correction) :

To : Marketing Director
Company ~~B&T~~ C-P-N
From : John Smith
SI, Deputy General Manager

SJ

Dear Sir,
Our company is the leader in the regional market in software development.
After more than 10 years in software development domain, we are in a position now to extend our activity to include the ~~installation and the configuration of computer networks~~ Marketing and sales of computers and printers.

We are expecting to ~~install~~ sell more than ~~25 local area network~~ 150 computers and 30 printers in the next 2 months and we suggest a transactional period of cooperation between our companies in aim to set the final conditions of corporation referring to our sales performance during this period.

Profiting from our position in the software market which shares the same customers of the ~~network~~ Hardware market we are able to achieve a high sales rate in a very limited period.
We are interested in your production line and we are interested to be your exclusive representative in the region. We also are able to setup a marketing department for your product line.
We are interested in the following items from your line:

Item	Quantity
~~- T-Junctions~~ Computer PC	~~5 000~~ 100
~~- Co-Axial cables~~ Printer	~~20 000~~ 50
~~- Terminators~~	~~1 000~~

Please send us your prices for these items and quantities for the first order.
Waiting for your reply, I remain

Sincerely Yours

John Smith
Deputy General Manager

Section 3
Annexe 2

Bibliographie de la section 3

[1] J.M. Christian Bastien et Dominique L. Scapin 1993, Critères ergonomiques pour l'évaluation d'interfaces utilisateurs, INIRIA Programme 3 - technical report N. 156.

[2] J.M. Chrisian Bastien, Dominique L. Scapin , Corine Leulier , Une comparaison des critères Ergonomiques et des principes de dialogue ISO 9241-10 dans une tâche d'évaluation d'interface. Rapport INIRIA.

[3] Christie, B.,& Gardiner, M. M. (1990). Evaluation of the human-computer interface. In J.R. Wilex & E.N. Corlett (Eds), Evaluation of human work: A practical ergonomics methodology (pp.271-320). London: taylor & Francis.

[4] Grislin, M., & kolski, C. (1996). Evaluation des interfaces homme- machine lors du développement des systèmes interactifs. Technique et science informatiques, 15,265-296.

[5] Howard, S., & Murray, D. M. (1987). A taxonomy of evaluation techniques for HCI, in proceedings of IFIP INTERACT'87: Human-Computer interaction (pp,453_459).

[6] Karat, J. (1988). Software evaluation methodologies. In M. Helandcr (Ed.) Handbook of human-computer interaction (pp.891_903). Amsterdam, the Netherlands: Elsovier Science Publishers.

[7] Lea, M. (1988). Evaluating User interface Designs. In T. Rubin (Ed.) User Interface Design for Computer Systems (pp.134_167). Chichester: Ellis- Horwood.

[8] Scapin, D.L. (1990) Des critères ergonomiques pour l'évaluation et la conception d'interfaces utilisateurs [Ergonomic criteria for the evaluation and design of user interfaces]. Actes du XXVI Congrès de la SELF. Montréal, Canada: institut de Recherche on Santé et Sécurité au Travail du Québec.

[9] Senach B. (1990). Evaluation ergonomique des interfaces homme- machine: une revue de la littérature (Rapport No.1180). Rocquencourt, France: institut National de Recherche en Automatique.

[10] Scnach B, (1993). L Evaluation Ergonomique des interfaces Homme-Machine. In J.C. Sperandio (Ed.)L Ergonomie dans la Conception des projets informatiques (pp.69-122). Toulouse, France.

[11] Sweeney M., & Dillon, A. (1987). Methodologies employed in the psychological evaluation of H.C.I. In proceedings of IFIP INTERACT 87: Human-Computer interaction (pp.367-373).

[12] Whitefield, A., Wilson, F., & Dowell, J. (1991). A framework for human factors evaluation. Behavior and information Technology.

Section 4

Etude temporelle de la multimodalité

Chapitre 8 : But et objectif de l'étude temporelle

Chapitre 9 : L'évaluation des interfaces homme-machine

Chapitre 10 : Etude temporelle expérimentale de la multimodalité

Annexe 1 : Résumé des principes d'opération du processeur humain

Annexe 2 : Les principes additionnels d'opération du processeur humain

Annexe 3 : Exemple de l'usage de la procédure de l'analyse de tâche par le modèle GOMS en utilisant NGOMSL

Annexe 4 : Bibliographie de la section 4

Chapitre 8

But et objectif de l'étude temporelle

Dans les chapitres précédents nous avons étudié l'usage et l'utilisabilité de la multimodalité d'un point de vue qualitatif. Dans cette section, l'étude aura une approche temporelle en se basant sur un modèle cognitif de l'être humain : GOMS développé par Card, Moran et Newell [5] pour l'étude des interfaces graphiques.

Ce modèle a été présenté à l'origine sans la modalité "parole". Cette section a pour but d'étendre le modèle à cette modalité puis à la multimodalité toute entière. Au préalable, dans le chapitre 9, nous rappelons le modèle "processeur humain" ou « *Human Processor* » présenté en 1983 par Card, Moran et Newell [5] dont le célèbre modèle GOMS était dérivé. Dans ce même chapitre, nous donnons aussi une méthodologie pratique d'utilisation du modèle GOMS. Cette méthodologie a été décrite par Kieras [6] pour donner un aspect plus pratique a GOMS, ce qui a conduit au langage de description des interfaces homme-machine NGOMSL " Natural GOMS Language".

Au chapitre 10, nous présentons une expérience sur une interface multimodale et nous étudions, avec NGOMSL, les résultats de cette expérience afin d'évaluer le gain ou le coût apporté par l'usage de la parole. Nous montrons également que la multimodalité a un coût dû au parallélisme. En complément, nous introduisons la parole et la multimodalité dans la notation NGOMSL. Enfin dans le chaitre 11 nous concluons l'ensemble de nos travaux.

Chapitre 9

L'évaluation des interfaces homme-machine

9-1- Introduction au modèle GOMS : "*le Processeur Humain*"

9-2- Langage de description des interfaces Homme-Machine NGOMSL

9-1- Introduction ou modèle GOMS : *"le Processeur Humain"*

Le concepteur de l'interface, placé en quelque sorte entre l'être humain et la machine, a besoin d'être capable d'analyser les tâches à accomplir par l'usager, et il a aussi besoin d'être capable de prédire et d'estimer *a priori* ses performances. Le modèle "Processeur Humain" donne quelques éléments pour ces questions.
Le modèle n'a pas comme but d'expliquer la cognition humaine, mais il dérive de modèles approximatifs faciles à utiliser. Le "cerveau" humain pouvant être décrit comme un système de traitement d'information, cette description ne concerne pas la réalité de ce qui se passe dans le cerveau humain mais elle concerne les grandes lignes du comportement humain.

Le modèle "le processeur humain" est décrit par:
1- Un ensemble de mémoires et de processeurs avec
2- Un ensemble de principes qui s'appellent "les principes d'opérations"

Il comprend trois sous-systèmes d'interaction:
1- Le système perceptuel,
2- Le système moteur,
3- Le système cognitif.

Chacun de ces trois systèmes a ses mémoires et processeurs.

Le système mémoires perceptuelles se compose de capteurs et de mémoires tampons associées à ces capteurs.
Le système cognitif reçoit des informations symboliques codées et les met dans sa mémoire à court-terme. Il utilise des informations stockées dans la mémoire à long-terme pour décider la façon de réagir. Le système moteur exécute l'action décidée par le système cognitif.

Le modèle "le processeur humain" supporte trois processeurs différents - processeur perceptuel, processeur moteur et processeur cognitif - qui réagissent de manière séquentielle ou en parallèle selon les tâches. Par exemple : dans le cas d'un usager qui recopie une lettre, chaque mot lu passe par le processeur perceptif puis par le processeur cognitif pour le reconnaître mais en même temps le processeur moteur donne des instructions pour taper le mot précédent.

Les mémoires et les processeurs du modèle sont décrits par des paramètres. Ceux qui concernent la mémoire sont :

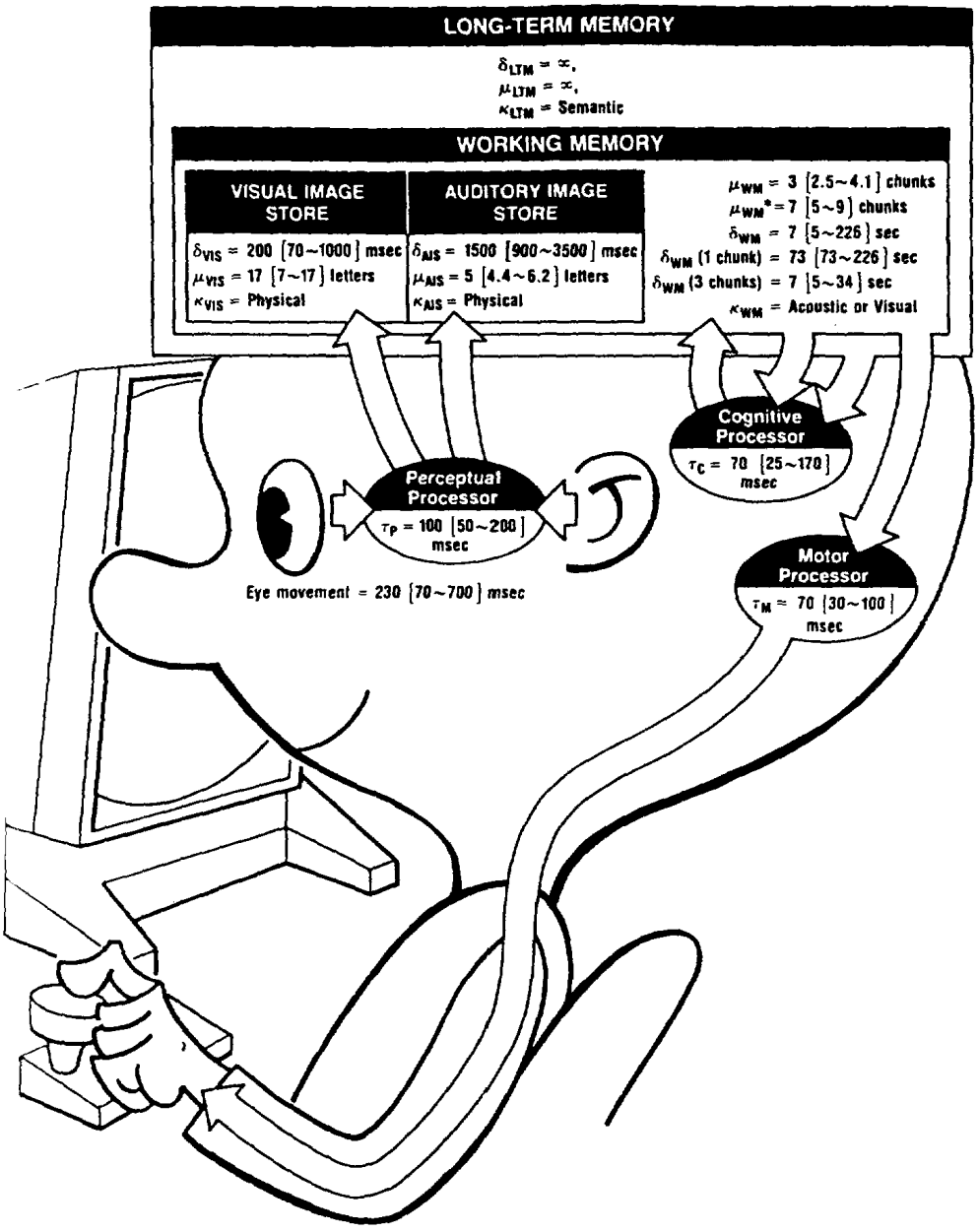
μ : la capacité de stockage,
δ : la constante d'oubli,

κ : le type du codage.

Ceux qui concernent le processeur sont :

τ : le temps d'un cycle (le temps pour traiter l'unité minimum d'information)

Il faut noter qu'il n'existe pas de paramètre spécifique pour le temps d'accès à la mémoire que l'on inclut généralement dans le temps de cycle du processeur.



Représentation schématique du "processeur humain" [5]

Le système perceptif : Le système perceptif traduit les informations sur le monde physique acquises par les capteurs du corps humain, en représentation interne.
Le modèle donne des estimations des paramètres, comme la durée du mouvement de l'œil, sous la forme suivante:

Mouvement de l'œil = 230 [70 ~ 700] m sec
- 230 : valeur typique
- 70 : limite minimale
- 700 : limite maximale

Les valeurs minimale et maximale représentent l'intervalle des valeurs possibles selon les différentes tâches, les différentes situations et les différents paradigmes ou méthodes de mesure. (pour la justification voir annexe 1)

Les mémoires perceptuelles :
Le codage : après la détection des informations visuelles ou acoustiques, il est effectué sur une représentation des ces informations puis stocké dans le *« visual image store »* ou *« auditory image store ».* Ces parties de la mémoire contiennent des informations codées physiquement. La durée du stockage est très courte.

L'interaction avec la mémoire du travail "*working memory*" :
La mémoire perceptuelle est reliée à la mémoire cognitive de travail. Les informations stockées dans la mémoire perceptuelle sont transférées dans la mémoire de travail et sont représentées symboliquement. Si la quantité d'information est importante, la mémoire de travail sera remplie avant le décodage de toutes les informations. Dans ce cas le processeur cognitif peut spécifier quelle partie des informations sera transférée.

Le temps de "Déclin" de la mémoire :
La notion de "demi-vie" est utilisée pour définir le temps d'oubli. Cet oubli est en fait progressif selon une courbe dite de "déclin".
La "demi-vie" est le temps après lequel la probabilité d'accès est inférieure à 50%. Le temps de "demi-vie" visuel est δ_{VIS} = 200 [90 ~ 1000] msec et le temps de "demi-vie" acoustique est δ_{AIS} = 1500 [900 - 3500] msec.

La capacité de la mémoire :
Il est très difficile de préciser la capacité de la mémoire perceptuelle visuelle ou acoustique mais une estimation pourrait être de :

μ_{VIS} = 17 [7 - 17] lettres
μ_{AIS} = 5 [4.4 - 6.2] lettres

Le processeur perceptif:

Le temps du cycle du processeur perceptif est autour de:

τ_p = 100 [50 - 200] msec.

Les perceptions qui prennent naissance dans le même cycle seront considérées comme combinées dans une perception unique si elles sont relativement similaires. En effet le temps du cycle du processeur perceptif est relié à l'intensité du stimulus : τ_p est plus court pour un stimulus plus intense (voir les principes d'opération l'annexe 1 de ce chapitre).

Le système moteur:

Le système moteur contrôle les muscles pour effectuer les actions. Le modèle "le processeur humain" a défini le mouvement comme une série de micro-mouvements chacun ayant besoin d'un temps d'exécution de durée τ_M.

$$\tau_M = 70 \; [30 \sim 100] \text{ msec.}$$

Le temps τ_M est identifié comme le temps du cycle du processeur moteur. Le cycle ou "feed-back" de l'action à la perception est suffisamment long (200 ~ 500 msec) ce qui nécessite pas que les actes comportementaux rapides comme la parole ou la frappe clavier doivent être considérés comme des "scripts" ou des instructions pré-programmées.

Le système cognitif :

Le système cognitif prend les entrées provenant du système perceptif pour donner les sorties convenables au système moteur. Les tâches exécutées par l'être humain sont complexes et ont besoin d'apprentissage, d'évaluation des effets des actions et des mécanismes généraux de résolution des problèmes : cela justifie la complexité des mémoires et du processeur cognitif du modèle.

Les mémoires cognitives :

Le modèle "Processeur humain" a deux mémoires dans son système cognitif : la mémoire de travail qui détient les informations pour l'utilisation immédiate et la mémoire à long-terme pour stoker les connaissances pour la future utilisation.

La mémoire de travail contient les résultats intermédiaires de raisonnement et les représentations produites par le système perceptuel. Dans la mémoire de travail toutes les opérations mentales travaillent sur des opérandes et produisent des sorties. La mémoire de travail est donc constituée de l'ensemble des registres du processeur cognitif qui sont des sous-ensembles des éléments de la mémoire à long-terme devenus actifs.

Le type du codage dans la mémoire de travail est visuel ou acoustique et non pas physique, ce qui diffère des informations stockées dans la mémoire de travail des informations non-symboliques et physiques de la mémoire perceptuelle qui sont affectées par les paramètres physiques des stimulus (comme l'intensité).

Les "*Chunk*" :

Les éléments activés de la mémoire à long-terme qui définissent la capacité de la mémoire de travail sont constitués des symboles qui s'appellent "*Chunk*" et qui peuvent être organisés dans d'autres *Chunks*.

p.e. chunk1 = (chunk2, chunk3, chunk4)
chunk4 = (chunk5, chunk6).

Le temps d'oubli :

Le temps d'oubli (demi-vie) varie dans un intervalle très large dû aux nombreux phénomènes d'interférence et de rafraîchissement des informations. Il est donc très difficile à évaluer.
On peut poser approximativement $\delta_{WM} = 7$ [5 - 256] secondes

En fait, la pente de la courbe d'oubli est très sensible au nombre des C*hunks* des éléments à rappeler, ce qui donne :

δ_{WM} (1 *Chunk*) = 73 [73 - 226] secondes
δ_{WM} (3 *Chunk*) = 7 [5 - 34] secondes.

La capacité pure du stockage de la mémoire de travail est de l'ordre de :

$$\mu_{WM} = 3 \text{ [2.5 ~ 4.2] chunks}$$

La capacité pure du stockage peut être augmenté avec l'aide de la mémoire à long- terme ce qui porte la capacité effective de la mémoire de travail à :

$$\mu^{*}_{WM} = 7 \text{ [5 ~ 9] chunks.}$$

La mémoire à long-terme :

La mémoire à long-terme retient les connaissances propres de l'individu. Elle est constituée de réseaux de chunks reliés qui sont accessibles à partir de la mémoire de travail de manière associative. Le contenu de la mémoire à long-terme est composé d'éléments portant sur : les effets des actions ou des raisonnements, les procédures opératoires et l'historique général.
Le temps d'oubli de la mémoire à long-terme est considéré dans le modèle processeur humain comme infini : $\delta_{LTM} = \infty$.

Il existe deux raisons pour lesquelles la recherche d'un *Chunk* dans la mémoire à long terme peut échouer :

1- L'association pour accéder a un *Chunk* (à partir de la mémoire du travail) est perdue, ou
2- Il existe plusieurs associations à des *Chunks* inférés dans la recherche du *Chunk* but.

Le codage des informations dans la mémoire à long-terme est sémantique. Il existe deux principes d'opération (p2 et p3) pour la mémoire qui a besoin de plusieurs opérations de recherche pour associer le nouvel élément avec des éléments déjà existants dans la mémoire. Cela rend l'écriture lente dans cette mémoire. Par contre un accès à cette mémoire est fait tous les 70 msec (le cycle du processeur cognitif), et on peut donc dire que la mémoire à long-terme opère comme un système "fast-read, slow-write".

Le processeur cognitif :

Le processeur cognitif a un cycle de reconnaissance-action ("*recognize-act*") qui est l'unité de base du traitement cognitif. A chaque cycle le contenu de la mémoire de travail initialise les actions dans la mémoire à long-terme (reconnaître) qui à son tour modifie le contenu de la mémoire de travail (acte) et prépare le lancement d'un nouveau cycle. Les plans, les procédures et les autres formes de comportements organisées sont constitués d'ensembles de cycles "reconnaître-agir".

Le temps du cycle est $\tau_c = 70$ [25 - 170] msec.

Le temps du cycle du processeur cognitif τ_c diminue quand un effort plus important est engagé.

τ_c diminue aussi avec le niveau de pratique acquis par apprentissage.

Le tableau suivant donne le temps de cycle du processeur cognitif pour plusieurs paradigmes expérimentaux:

Rate at wich an item can be mached against working memory

Digits	33 [27 – 39] msec/item
Colors	38 msec/item
Letters	40 [24 – 65] msec/item
Words	47 [36 – 52] msec/item
Geometric shapes	50 msec/item
Random forms	68 [42 – 93] msec/item
Nonsense syllables	73 [27 – 93] msec/item

Rate at which four or fewer objects can be counted

Dot patterns	46 msec/item
3-D shapes	94 [40 – 172] msec/item
Perceptual judgement	106 [85-169] msec/inspection
Choice reaction time	92 msec/inspection 153 msec/bit
Silent counting rate	167 msec/digit

Le système cognitif travaille en parallèle dans sa phase de reconnaissance et en séquentiel dans sa phase d'action. A ce moment le système cognitif peut prendre conscience de plusieurs choses en même temps mais il n'est pas capable de les faire en même temps.

9-2- Langage de description des interfaces Homme-Machine NGOMSL

Le modèle GOMS :

Le modèle GOMS est fondé sur le principe qui dit que pour prévoir le comportement de l'utilisateur on doit analyser la tâche pour définir le but de l'utilisateur et les contraintes de la tâche. Avec le modèle "le processeur humain" on peut détailler le comportement d'un utilisateur à partir des opérateurs de traitement d'information et du comportement de l'utilisateur décrit au moyen d'une séquence d'opérations élémentaires. Le temps d'exécution d'une tâche est alors l'accumulation des temps de chaque opération mise en jeu.
Le Modèle GOMS considère que la structure cognitive de l'utilisateur est constituée de quatre éléments : un ensemble de buts, un ensemble d'opérateurs, un ensemble de méthodes pour atteindre ces buts et un ensemble de règles de sélection pour choisir les méthodes convenables pour atteindre le but en question. Une méthode est constituée de plusieurs opérateurs. Les détails du modèle "*GOMS*" sont présentés dans [5] ainsi que le modèle "le processeur humain". Le modèle GOMS a été utilisé dans plusieurs travaux d'évaluation des interfaces homme- machine [4].
La présentation de GOMS décrite par Card, Moran et Newell [5] ne détaille pas la notation, elle parait difficile à utiliser et n'a pas des bases bien définies pour le calcul du temps d'apprentissage explicite [6][9].

Pour dépasser ces inconvénients, Keiras [6] a proposé un langage de description de tâches "*Task Analysis*" qui s'appelle ***NGOMSL*** "*Natural GOMS Language*". Basé sur le modèle GOMS et ayant la structure d'un langage de programmation. NGOMSL est facile à utiliser et peut être exploité pour l'évaluation d'une interface, pour prévoir les performances humaines et pour réviser la conception et la documentation d'une interface Homme-Machine.

Les paragraphes suivants présentent rapidement l'analyse basée sur le modèle GOMS avec le langage NGOMS. La présentation détaillée se trouve dans [5].

L'analyse GOMS est un modèle de la connaissance que l'utilisateur doit avoir pour accomplir des tâches sur une machine. Elle est aussi une représentation de connaissances "comment faire" que le système doit avoir pour pouvoir accomplir la tâche en question. Le but de l'analyse de tâches GOMS est de décrire formellement les buts, les opérateurs, les méthodes et les règles de sélection.

Les buts :

Les buts peuvent être organisés d'une façon hiérarchique, chaque but est composé de plusieurs sous-buts. Pour atteindre un but il faut d'abord atteindre les sous-buts qui le composent. La description d'un but est une paire action-objet et elle a la forme < verbe nom> comme "effacer mot".

Les opérateurs :

Les opérateurs sont des actions que l'utilisateur exécute. Les opérateurs ont la même forme que les buts (action- objet), mais la différence entre les deux est que les buts sont à atteindre tandis que les opérateurs sont à exécuter. Les buts sont atteints par l'exécution de plusieurs opérateurs. Les opérateurs sont décomposés en méthodes jusqu'aux niveaux les plus bas de l'analyse. Cette procédure est récursive jusqu'à ce qu'on arrive à des opérateurs primitifs qu'on ne peut plus décomposer.

Les actions par lesquelles l'utilisateur échange des informations avec l'environnement s'appellent des "opérateurs externes". P.e. lire un texte de l'écran (perceptuelle), enfoncer un bouton (moteur) ou tourner une page d'un manuscrit, etc.

Les actions internes faites par l'utilisateur sont des opérations mentales p.e. faire un choix, stocker ou chercher des informations dans la mémoire de travail ou la mémoire à long-terme.

La notation NGOMSL pour les primitives mentales est :

→ **Accomplish the goal of <goal description>**
(analogue à un appel de procédure " CALL statement")

→ **Report goal accomplished**
(analogue à "RETURN statement")

→ **Decide: If <operator ---> Then <operator>**
Decide: If <operator ---> Then <operator> Else <operator>

(analogue à si --- alors --- sinon ---)

→ **Go to step <number>**
(analogue à "GOTO statement"). Cette instruction est utilisée normalement avec l'opérateur "Decide".

→ Les opérateurs de stockage et à la mémoire:
Recall that <WM- object- description>
Retain that <WM- object- description>
Forget that <WM- object- description>
Retrieve- LTM that
<LTM- object- description>

Recall: chercher de la mémoire
Retain: stoker dans la mémoire
Forget: éliminer de la mémoire

On remarque qu'il n'existe pas un opérateur unique qui concerne la mémoire à long-terme parce que l'apprentissage et l'élimination (oublier) des informations de la mémoire à long-terme ne sont pas engagées dans les tâches à modéliser.

Les opérateurs externes primitifs :
En faisant l'analyse, on peut définir des primitives basées sur des actions élémentaires nécessaires pour le système.

P.e.: Home- hand to mouse
Press- key <key name>
Type- in <string of characters>
Move- cursor to <target coordinates>
Find- cursor- is- at <returned cursor coordinates>
Find- menu item <menu- item- description>.

Les opérateurs mentaux définis par l'analyste :
En faisant l'analyse, on trouve des cas où l'utilisateur est engagé dans des processus psychologiques très complexes qui ne sont pas représentables dans le modèle GOMS. La plupart du temps ces processus n'ont pas d'influence sur la conception de l'interface et on peut les négliger. P.e.
Verify- result: vérifier si le résultat est bon.
Get- next- edit- location: chercher l'endroit où il faut modifier.

Les méthodes :
Une méthode est une séquence d'étapes pour atteindre un but.

L'étape dans une méthode est un opérateur externe ou un ensemble d'opérateurs mentaux qui sert à atteindre un sous-but. En réalité, l'analyse des interfaces consiste essentiellement à spécifier les étapes que l'utilisateur doit accomplir pour atteindre son but. La description des méthodes est le cœur de l'analyse des tâches.
La forme d'une méthode est:

Method to accomplish goal of <goal description>
Step1. <operator>…
Step2. <operator>…
…
Step3. Report goal accomplished.

Exemple:
la méthode pour déplacer un fichier dans un système Windows 95:

Step1. Select the icon for the file
Step2. Drag the icon to the destination icon
Step3. Report goal accomplished

Les méthodes peuvent appeler des sous-méthodes pour atteindre des sous-buts :

Method to accomplish goal of <goal description>
Step1. <operator>
Step2. <operator>
…
Step k. Accomplish the goal of <sub-goal description>
…
Step n. Report goal accomplished.

Method to accomplish goal of <sub-goal description>
Step1. <operator>
Step2. <operator>
…
Step k. Accomplish the goal of <sub- sub-goal description>
…
Step n. Report goal accomplished.
…

La procédure d'estimation utilisée fait référence au nombre de "Statements" de NGOMSL. Chacun de ces "Statements" correspond à une règle de production. Le "Statement" d'une méthode compte pour un seul "Statement", et chaque

étape compte pour un "Statement" sans considérer le nombre des opérateurs dans l'étape.

L'opérateur "Decide " avec un "Else" compte pour deux "Statements" parce que deux règles de productions sont engagées.

Les Règles de Production

Les règles de production ont un rôle de contrôle de la méthode adéquate pour atteindre le but puisque pour atteindre un but il existe plusieurs méthodes. Le but général doit être décomposé en sous-buts (un pour chaque méthode) et des conditions d'exclusion mutuelle doivent spécifier quelle méthode sera choisie dans un contexte donné.

Dans la notation NGOMSL les règles de sélection sont regroupées en paquets. Chaque paquet est associé à un but et constitué de plusieurs règles "*if- then*" pour choisir le but spécifique à atteindre.

La forme d'une règle de sélection est:

Selection rule set for goal of <general goal description>

If <condition> Then
accomplish goal of <specific goal description>.
If <condition> Then
accomplish goal of <specific goal description>.
…
Report goal accomplished.

Chaque condition consiste en un ou plusieurs opérateurs qui testent la mémoire de travail, la description de la tâche ou la situation perceptuelle externe. Ces opérateurs ne peuvent pas être des opérateurs actionnels comme "enfoncer un bouton". L'opérateur "*Decide*" n'est pas utilisé dans les règles de sélection.
Le "*Selection Rule Set*" et "*Report Goal Accomplished*" compte pour un "*statement*" chacun. Chaque "*if- then*" compte aussi pour un "*statement*". L'ordre des "*if - then*" n'a pas d'importance mais il faut qu'ils soient écrits de façon qu'une seule condition soit vraie en même temps. Dès qu'un but spécifique est atteint par un "*if- then*", le but général est aussi atteint, le résultat est noté par "*Report goal accomplished*".

La description des tâches et les instances des tâches:

La description des tâches décrit la tâche en terme de but à atteindre et de "liste de paramètres" des méthodes qui accomplissent la tâche. Une instance de tâche

est une description de tâche qui contient des valeurs pour tous les attributs de la description de la tâche.

La procédure de construction d'un modèle GOMS:

La construction d'un modèle GOMS est une procédure « descendante » (top-down) et « largeur d'abord » (breadth first). L'analyse est faite à partir du but général en descendant vers les sous-buts et jusqu'aux opérateurs primitifs à la fin de la procédure. Tous les buts du même niveau sont traités avant d'aller au plus bas niveau.

L'utilisation de l'analyse de GOMS (notation NGOMSL) :

Cette analyse est utile pour évaluer les interfaces des systèmes existants, pour l'évaluation des systèmes dans l'étape de développement et pour l'évaluation des interfaces pendant la conception. Les principaux points de l'utilité de l'analyse de tâche "GOMS" sont pour :

→ L'évaluation qualitative de la conception, (en répondant aux questions comme) :

- Est ce que l'interface est naturelle ? L'utilisateur est-il obligé d'apprendre une nouvelle façon de penser pour s'adapter à l'interface ?

("*Naturalness of the design*")

- Est-ce qu'il existe des méthodes pour atteindre chaque but ou sous-but de l'interface ?
("*Completeness of the design*").

- Est-ce qu'il existe plus d'une méthode pour atteindre un but ?
Si oui quelques méthodes sont-elles nécessaires ?
("*Cleanliness of the design*").

- Est-ce que les buts similaires sont atteints par des méthodes similaires ?
("*Consistency of the design*").

- Est-ce que les buts sont atteints par des méthodes concises et rapides ?
("*Efficiency of the design*").

→ Prédire les performances humaines pour une interface donnée

→ L'estimation du temps d'apprentissage: (Keiras [6])

"***Learning Time = (30 - 60) minutes + 30 sec per Number of NGOMSL statements***".

→ L'estimation du temps d'exécution: (Keiras [6])

Execution Time =
NGOMSL statement time
+ Primitive External Operator Time
+ Analyst- defined Mental Operator Time
+ System Response Time

NGOMSL Statement Time =
Number of NGOMSL statements executed * 0.1 secs

Primitive External Operator Time =
Total of times for external operators defined by the analyst

System Response Time = Total time when user is idle

0.28sec for a mouse button press or release
0.1 sec for a mouse button press or release
1.1 sec for (average)for a mouse move
0.4 sec to home the hand to a keyboard or a mouse

Le modèle GOMS peut être utilisé pour estimer la charge mentale d'une tâche et pour réviser une conception. Il sert aussi pour la documention du système (Keiras [6]).

Chapitre 10

Etude temporelle expérimentale de la multimodalité

1- L'étude expérimentale – but, objectif et description générale

2- Description de l'expérience selon la notation NGOMSL

3- L'expérience : les mesures et les résultats

4- Observations et commentaires

5 - Conclusion

1- L'étude expérimentale – but, objectif et description générale :

Après avoir présenté le modèle GOMS et ses aspect pratiques dans le chapitre précédent, nous allons utiliser les principes de ce modèle pour l'étude temporelle de la multimodalité.

Le but de cette étude est de prouver que la multimodalité a (ou n'a pas) de coût supplémentaire par rapport à l'utilisation de plusieurs modes d'interaction pris séparément.

Dans le cadre de cette étude, nous avons effectué une expérience sur quelques utilisateurs (5 utilisateurs) qui sont entraînés à l'interaction multimodale et qui maîtrisent bien le logiciel de reconnaissance de la parole utilisé et les commandes admissibles. Les sujets sont aussi des experts de l'environnement de travail dans lequel l'expérience est effectuée. Le logiciel de reconnaissance des commandes vocales est le logiciel *«Microsoft Voice»* qui a été décrit dans le chapitre 7.

L'expérience était filmée par une camera vidéo et elle était analysée ulterieuement sur l'écren d'un téléviseur et à l'aide d' un chronomètre manuel de précision de 10 ms. Les mêmes mesures de temps étaient effectuées plusieurs fois et pour plusieurs personnes, et nous avons considéré les résultats de l'utilisateur qui avait la meilleure performance en multimodalité. Des méthodes qui servent à assurer une bonne estimation du temps de chaque sous-tâche étaient utilisées aussi afin d'avoir les résultats avec la meilleure précision possible. (p.e. pour avoir le temps moyen d'une sous-tâche quelconque nous avons effectué la tâche qui contient cette sous-tâche deux fois: une fois avec cette sous-tâche et une autre fois sans cette sous-tâche et nous avons comparé les deux résultats avec le temps mesuré pour cette sous-tâche afin de nous assurer de nos mesures).

Le but de la tâche à accomplir dans l'expérience est de déplacer une ligne de texte en passant par les étapes suivantes :

1- Aller vers la ligne à déplacer (aller(x))
2- Sélectionner la ligne à déplacer (sélectionner (x))
3- Couper la sélection (couper(x))
4- Aller à la destination (aller(y))
5- Coller la sélection (coller(x)(y))

Ce but pouvait être atteint de plusieurs façons :

- En utilisant le mode geste (clavier et souris)

- En utilisant le mode parole
- En utilisant la multimodalité (mode parole + mode geste)

Le scénario de l'exécution est imposé au sujet pour éliminer les effets de la planification et les sujets avaient le temps pour s'entraîner avant l'expérience pour obtenir la meilleur performance en effectuant la tâche en question.

Les scénarios de l'exécution sont les suivants :

a- Effectuer la tâche en utilisant le mode geste :

Aller (x)	➔	souris
Sélectionner (x)	➔	clavier (Shift+End)
Couper (x)	➔	clavier (Ctrl+x)
Aller (y)	➔	clavier (pagedown – pagedown)
Coller (x)(y)	➔	clavier (Ctrl+v)

Dans le scénario ci-dessus, on peut constater que nous avons forcé le sujet à changer de média pour effectuer les sous-tâches. Ce passage entre les differents médias, rend notamment plus facile la détection du début et de la fin de chaque sous tâche.

b- Effectuer la tâche en utilisant le mode parole :

Aller (x)	➔	commande (*Page Down*) (1 fois)
Sélectionner (x)	➔	commande (*Select Line*)
Couper (x)	➔	commande (*Cut Selection*)
Aller (y)	➔	commande (*Page Down*) (2 fois)
Coller (x)(y)	➔	commande (*Past Selection*)

c- Effectuer la tâche en multimodalité :

Aller (x)	➔	souris	geste
Sélectionner (x)	➔	commande (*Select Line*)	parole
Couper (x)	➔	commande (*Cut Selection*)	parole
Aller (y)	➔	clavier (2 fois PAGE DOWN)	geste
Coller (x)(y)	➔	commande (*Past Selection*)	parole

A partir des mesures de temps d'exécution de ces scénarios et en faisant la mesure en utilisant la notation NGOMSL nous allons étudier maintenant le coût de la multimodalité.

2- Description de l'expérience selon la notation NGOMSL :

Les scénarios précédents sont décrits ci-dessous selon la notation NGOMSL :

Method A to accomplish goal of moving text
Step 1 *Accomplish goal of selecting cutting point*
Step 2 *Accomplish goal of cutting text (call method B)*
Step 3 *Accomplish goal of pasting text (call method C)*
Step 4 *Report goal accomplished*

Method B to accomplish goal of cutting text
Step 1 *Accomplish goal of selecting text*
Step 2 *Accomplish goal of issuing CUT command*
Step 3 *Report goal accomplished*

Method C to accomplish goal of pasting text
Step 1 *Accomplish goal of selecting insertion point*
Step 2 *Accomplish goal of issuing PASTE command*
Step 3 *Report goal accomplished*

Les tableaux suivants montrent les différentes étapes et leurs opérateurs externes ainsi que le temps associé **à chaque étape :**

a- Le scénario en utilisant le mode geste :

NGOMSL statement	**Primitive External Operator**	**Time (in sec) NGOMSL estimations**
Accomplish goal of selecting cutting point (method A)		0.1
	Home the hand to the mouse	0.4
	Mouse move	1.1
	Mouse button press	0.1
Accomplish goal of selecting text (method B)		0.1
	Home the hand to the keyboard	0.4
	Key stroke	0.28
	Key stroke	0.28
Accomplish goal of issuing CUT command (method B)		0.1
	Home the hand to the keyboard	0.4
	Key stroke	0.28
	Key stroke	0.28
Accomplish goal of selecting insertion point(method C)		0.1
	Home the hand to the keyboard	0.4
	Key stroke	0.28
	Key stroke	0.28
Accomplish goal of issuing PASTE command (method C)		0.1
	Home the hand to the keyboard	0.4
	Key stroke	0.28
	Key stroke	0.28
Total		**5.94**

b- Le scénario en utilisant le mode parole :

NGOMSL statement	Primitive External Operator	Time (in sec)
Accomplish goal of selecting cutting point (method A)		0.1
	Command: PAGE DOWN	No estimation in NGOMSL
Accomplish goal of selecting text (method B)		0.1
	Command : SELECT LINE	No estimation in NGOMSL
Accomplish goal of issuing CUT command (method B)		0.1
	Command : CUT SELECTION	No estimation in NGOMSL
Accomplish goal of selecting insertion point(method C)		0.1
	Command : PAGE DOWN	No estimation in NGOMSL
	Command : PAGE DOWN	No estimation in NGOMSL
Accomplish goal of issuing PASTE command (method C)		0.1
	Command : PASTE SELECTION	No estimation in NGOMSL

c- Le scénario en utilisant la multimodalité :

NGOMSL statement	Primitive External Operator	Time (in sec)
Accomplish goal of selecting cutting point (method A)		0.1
	Home the hand to the mouse	0.4
	Mouse move	1.1
	Mouse button press	0.1
Accomplish goal of selecting text (method B)		0.1
	Command : SELECT LINE	No estimation
Accomplish goal of issuing CUT command (method B)		0.1
	Command : CUT SELECTION	No estimation
Accomplish goal of selecting insertion point(method C)		0.1
	Home the hand to the keyboard	0.4
	Key stroke	0.28
	Key stroke	0.28
Accomplish goal of issuing PASTE command (method C)		0.1
	Command : PASTE SELECTION	No estimation

La notation NGOMSL ne donne pas d'estimations sur le temps d'exécution des commandes vocales. NGOMSL ne prend pas les effets de la multimodalité en considération. Dans le paragraphe qui suit nous allons étendre GOMS et nous allons estimer (pour notre expérience) le temps d'exécution de chaque commande vocale par rapport au temps de réaction du système.

3- L'expérience : les mesures et les résultats

Les scénarios de l'expérience que nous avons présentés ci-dessus sont effectués par les sujets de test. Nous avons obtenu les résultats suivants :

a- en utilisant le mode geste : (les temps présentés sont les temps obtenus par l'utilisateur qui a eu les meilleurs temps en multimodalité).

Opération	Temps d'exécution sec
Aller(x) {souris}	1.5
Sélectionner (x) {clavier}	0.9
Couper(x) {clavier}	0.9
Aller (y) {clavier 2fois Page Down}	0.8
Coller (x)(y) {clavier}	0.9
	5.0

b- en utilisant le mode parole :

Opération	Temps d'exécution Sec
Aller(x) {PAGE DOWN}	1.5
Sélectionner (x) {SELECT LINE}	1.7
Couper(x) {CUT SELECTION}	1.9
Aller (y) {PAGE DOWN – PAGE DOWN }	3.4
Coller (x)(y) {PASTE SELECTION}	1.8
	10.3

Il faut noter dans ce cas que le temps de chaque commande est le temps de prononciation plus le temps de réponse nécessaire au système plus le temps d'attente entre deux commandes.

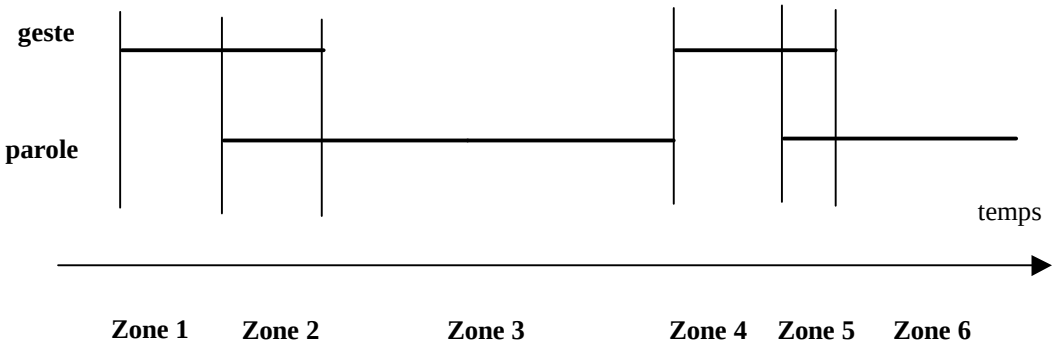
Une mesure fine du temps de prononciation des commandes est faite et le temps moyen pour chaque commande utilisée est :

Commande	Temps de pronociation
Page Down	740 msec
Select Line	950 msec
Past Selection	1050 msec
Cut Selection	130 mesc

c- en utilisant la multimodalité :

Le temps total pour accomplir la tâche est entre 6.8 et 8 secondes (pour l'utilisateur qui a eu le meilleur temps) et le meilleur temps enregistré est de 6.8 secondes.

Le scénario idéal pour exploiter les avantages du parallélisme offerts par le système est le suivant (c'est le scénario qui a obtenu le meilleur temps) :



Selon le schéma précédent la zone 1 est la zone du geste (place le curseur avec la souris). La zone 2 est la zone de la prononciation de la commande (Select Line). La zone 3 est la zone pendant laquelle le système reconnaît la commande (Select Line) et l'exécute et l'utilisateur prononce la commande (Cut Selection) et le système la reconnaît et l'exécute. La zone 4 est la zone ou l'utilisateur tape "2" sur la touche (Page Down). La zone 5 est la zone dans laquelle l'utilisateur peut prononcer la commande (Paste Selection) avant que la sous tâche de mettre le curseur dans sa position soit achevée. Dans la zone 6, le système reconnaît et exécute la commande (Paste Selection).

4- Observations et commentaires

Théoriquement on peut diminuer le temps d'exécution en cas d'usage de la multimodalité en éliminant le temps de prononciation (zone 2 et zone 5) qui est égale à 2 sec (950msec+1050 msec).

Le tableau suivant montre le temps d'exécution théorique estimé en éliminant les effets de la multimodalité (les chiffres donnés sont les valeurs des expériences en monomodalité geste ou parole):

Opération	Temps d'exécution prédit (sec)
Aller(x) {souris}	1.5
Sélectionner (x) {SELECT LINE}	1.7
Couper(x) {CUT SELECTION}	1.9
Aller (y) {clavier: 2 fois PAGE DOWN }	0.8
Coller (x)(y) {PASTE SELECTION}	1.8
	7.7

Si l'utilisation de la multimodalité n'a pas de coût supplémentaire, l'utilisation parallèle doit améliorer le temps d'exécution de cette tâche pour atteindre théoriquement :

7.7 sec – 2 sec (dû au gain du parallélisme) = 5.7 msec

Les mesures de l'expérience indiquent au meilleur cas le temps de 6.8 sec. (les mesures du temps des autres utilisateurs ont conduit à des résultats relativement identiques).

Ce résultat nous amène à conclure que la multimodalité a un certain coût dû d'un côté au parallélisme et d'un autre côté au passage d'un mode à l'autre (rémanence modale).

5- Conclusion

Le modèle GOMS et la notation NGOMSL ont dû être étendus pour tenir compte d'opérateurs externes liés à la parole (et plus généralement à la multimodalité). Les mesures nécessaires pour cette extension devront être faites par des laboratoires spécialisés et qui ont des outils de mesures très précis. Cette extension pourra alors servir à l'évaluation des interfaces homme-machine multimodales.

Nous avons montré cependant, avec les moyens dont nous disposions que la multimodalité à un double coût : celui de la gestion du parallélisme et celui de la rémanence d'un mode (coût de passage d'un mode à un autre).

Cette étude a porté sur plusieurs utilisateurs dans le but de généraliser les résultats et sans tenir compte des caractéristiques de chaque utilisateur. La relation entre le coût de la multimodalité et les caractéristiques des utilisateurs pourrait aussi être le sujet d'une future étude.

Section 4
Annexe 1

Résumé des principes d'opération du processeur humain

Résumé présenté en anglais pour conserver la notation et les expressions utilisées par les auteurs Card, Moran et Newell [5]:

1- **Duration of Saccadic Eye Movements**

Eye movement = 230 [70~700]msec.

Actual saccadic eye- movement times (travel + fixation time) can vary considerably depending on the task and the skill of the observer. Russo (1978, Table2, p.94) lists 70 msec as the minimum time and 230 msec as a typical time. The largest time given by Busswell (1922, p.31) for eye movements in reading is 660 msec (for first- grade children), which we round to 700 msec.

2- **Decay Half- Life of Visual Image Store**

δ_{VIS} = 200 [90~1000] mesc.

A least-squares fit to data estimated from figures appearing in Sperling (1960) and Averbach and Corielll (1961) yields the following facts. The half-life of the letters in excess of the memory span that subjects could report in the partial report condition of Sperling's (1960) experiment was 621 msec (9-letter stimulus) and 215 msec (12-letter stimulus). Averbach and Coriell's (1961) experiment gives a half- life of 92 msec (16-letter stimulus). The typical value for δ_{VIS} has been set at 200 msec, representing the middle of these. The lower and upper bounds for δ_{VIS} are set at rounded-off values reflecting the fastest subject in the condition with the shortest half-life. and the slowest subject in the condition with the longest half-life. The shortest half-life in these experiments was 93 msec for Averbach and Coriell's subject GM (16-letter condition); the longest half-life was 940 msec for Sperling's Subject ROR (9-letter condition). It is possible to have the average half-life be 92 msec, shorter than the half-life of any subject, because this average is computed by first taking the mean of each

point across subjects, then computing the slope of the best least-square fitting line in semi coordinates.

3- Decay Half-Life of Auditory Image Store

$$\delta_{AIS} = 1500\ [900\sim3500]\ \text{msec.}$$

The half-life of the letters in excess of the memory span that subjects could report in the partial report condition of Darwin, Turvey, and Crowder's (1972) experiment was 1540 msec, which we have rounded to $\delta_{AIS} = 1500$ msec. The difference in decay half-life as a function of letter order in their experiment (963 msec for the third letter, 3466 msec for the first letter) has been rounded to give lower and upper bounds of 900 and 3500. Other techniques have been used to obtain values for the "decay time" of the auditory image store. For example, use of a masking technique gives estimates of around 250 msec full decay (Massaro, 1970), but these experiments have been criticized by Klatzky (1980, p.42) because they may only measure the time necessary to transmit categorical information to working memory. On the other end, experiments that measure the delay at which there is still some facilitation of the identification of a noisy signal (Crossman, 1958; Guttman & Julesz, 1963) give very wide full-decay estimates: from 1000 msec to 15 minutes!

4- Capacity of Visual Image Store

$$\mu_{VIS} = 17\ [7\sim17]\ \text{letters.}$$

Sperling (1963, p.22) estimates the capacity of the visual image store in terms of the number of letters available, at least 17 letters and possibly more. The fewest number of letters available for any subject immediately after stimulus presentation in the nine-letter condition (Sperling, 1960) was 7.4 letters for Subject NJ.

5- Capacity of Audiotory Image Store

$$\mu_{AIS} = 5\ [4.4\sim6.2]\ \text{letters.}$$

Range is from the number of letters or numbers that could be reported by Darwin, Turvey, and crowder's (1972) subjects in an experiment in which they had to give the trio of letters coming from one of three directions (indicated by a visual cue shortly after the end of the end sounds). The lowest value, 4.4 letters, is for accuracy of recalling second letter of triple when subjects had to name all items coming from a certain direction (Fig.1, p.259). The highest number, 6.2

letters, is for recall by category when no location was required (Fig. 2 (B), p.262).

6- **Cycle Time of the Perceptual Processor**

$$\tau_p = 100\ [50 \sim 200]\ \text{msec.}$$

The source of the range is the review by Harter (1967), who also discusses the suggestion that the cycle time can be identified with the 77-125-msec alpha period in the brain.

7- **Cycle Time of Motor Processor**

$$\tau_M = 70\ [30 \sim 100]\ \text{msec.}$$

The limit of repetitive movement of the hand, foot, or tongue is about 10 movements per second (Fitts & Posner, 1967, p.18). Chapanis, garner, and Morgan (1949, p. 284) cite tapping rates of 8-13 taps per second (38-62 movements per second, assuming two movements per tap). Fox and Stansfield (1964) cite figures of 130 msec per tap = 65 msec per movement. Repetition of the same key in Kinkead's (1975) data (Figure 45.11) averages to 180 msec per keystroke = 90 msec per movement. We summarize these as 70 [30~100] msec per movement.

8- **Decay Half-Life of Working Memory**

$$\delta_{WM} = 7\ [5 \sim 226]\ \text{second}$$

$$\delta_{WM}(1\ \text{chunk}) = 73\ [73 \sim 226]\ \text{second}$$

$$\delta_{WM}(3\ \text{chunk}) = 7\ [5 \sim 34]\ \text{second}$$

For three chunks, Peterson and Peterson's (1959) data (Figure 45.5) give a half-life of about 5 seconds. Murdock's data (Murdock, 1961) in Figure 45.5 give a half-life of about 7 seconds for three words and also 9 seconds for three consonants. On the other hand, Melton's (1963) data give a much longer half-life of 34 seconds. For one chunk, Murdock's data in Figure 45.5 and Melton's (1963) give half-lives of 73 and 226 seconds, repectively.

9- **Pure Capacity of Working Memory**

$$\mu_{WM} = 3\ [2.5 \sim 4.1]\ \text{chunks.}$$

Crowder (1976) reviews several methods. Estimates are Waugh and Norman (1965) method, 2.5 items; Raymond (1969) method, 2.5 items; Murdock (1960, 1967) method, 3.2- 4.1 items; Tulving and Colatala (1970) method, 3.3- 3.6 items. See also Glanzer and Razel (1974).

10- **Cycle Time of Cognitive Processor**

$$\tau_C = 70\ [25\sim170]\ \text{msec.}$$

On the fast end, memory scanning rates go down to 25 msec per item (Sternberg, 1975, p. 225, Figs. 8 and 9, lower error bar for LETTERS). Michon (1978, p.93) summarizes the search for the "time quantum" as converging on 20- 30 msec. On the slow end, silent counting, which takes about 167 msec per item (Landauer, 1962), has sometimes been taken as a minimum cognitive task. It has sometimes been argued (Hick, 1952) that the subject in a choice reaction time experiment makes one choice for each bit in the set of alternatives, in which case a typical value would be 153 msec/bit (Figure 45.16) Welford (1973) has proposed a theory of choice reaction in which the subject makes a series of choices, each taking 92 msec. Blumenthal (1977) reviews an impressively large number of cognitive phenomena with time constraints in the 0.1- second range. The typical value has been set at 70 msec, about the median of the values in table 45.2.

11- **Fitt's Law Slope Constant**

$$I_M = 100\ [50\sim120]\ \text{msec/bit}$$

For single, discrete, subject-paced movements, the constant is a little less than $I_M = 100$ msec/bit and closer to the 50~68 msec/bit value cited above for other experimental methods and for our norminal calculation. Fitts and Peterson (1964) get 70~75 msec/bit. Fitts and Radford (1966) get maximum rates of 85 msec/bit (11.7 bits/sec) in a pointing experiment- paced tasks, such as alternately touching two targets with a stylus or pursuit tracking , the constant is a little above I_M =100 msec/bit. Elkind and Sprague (1961) get maximum rattes of 135 msec/bit (7.4 bits/sec) for a pursuit- tracking task. Fitts's orginal dotting experiment (Figure 45.8) gives 118 msec/bit using Eq. (3). Welford's (1968) study using Eq. (3) and the actual distance between the dots gives 120 msec/bit.

The references used by the authors in the original publication:

Averbach, E., & Coriell, A.S. Short-term memory in vision. Bell System Technical Journal, 1961,40,309-328.

Blumenthal, A.L. The process of cognition. Englewood Cliffs, N.J.: Prentice-Hall, 1977.

Busswell, G.T. Fundamental reading habits: A study of their development Education Monographs (Supplement), 1922, 21.

Chapanis, A., Garner, W. R., & Morgan, C. T. Applied experimental psychology: Human factors in engineering design. New York: Wiley, 1949.

Crossman, E. R. F. W., & Goodeve, P.J. Feedback control of hand movements and Fitts, law. Paper presented at the meeting of the Experimental Psychology Society, Oxford, July 1963. Quarterly Journal of Experimental Psychology, 1983, 35A, 251-278

Crowder, R. G. Princinciples of learning and memory. Hillsdale, N.J.: Erlbaum, 1976.

Darwin, C. J., Turvey, M.T., & Crowder, R. G. An auditory analogue of the Sperling partial report procedure: Evidence for brief auditory storage. Cognitive Psychology, 1972, 3, 255-267.

Elkind, J.I., & Sprague, L. T. Transmission of information in simple manual control systems. IEEE Transaction on Human Factors in Electronics, 1961, HFE-2,58-60.

Fitts, P. M., & Peterson, J. R. Information capacity of discrete motor responses. Journal of Experimental Psychology, 1964, 67, 103-112.

Fitts, P.M., & Posner, M. I. Human performance. Belmont, Cal.: Brooks Cole, 1967.

Fitts, P. M., & Radford, B. Information capacity of discrete motor responses under different cognitive sets. Journal of Experimental Psychology, 1966, 71, 475-482.

Fox, J.G., & Stansfield, R. G. Diagram keying times for typists. Ergonomics, 1964, 7, 317-320.

Glanzer, M., & Razel, M. The size of the unit in short-term storage. Journal of Verbal Learning and Verbal Behavior, 1974, 13, 114-131.

Guttman, N., & Julesz, B. Lower limits of auditory periodicity analysis. Journal of the Acoustical Society of America, 1963, 35, 610.

Harter, M. R. Excitability and cortical scanning: A review of two hypotheses of central intermittency in perception. Psychological Bulletin, 1967, 68, 47-58.

Hick, W. E. ON the rate of gain of information. Quarterly Journal of Experimental Psychology, 1952, 4, 11-26.

Kinkead, R. Typing speed, keying rates, and optimal keyboard layouts. Proceedings of the Nineteenth Annual Meeting of the Human Factors Society, 1975.

Klatzky, R. L. Human memory: Structures and processes (2nd ed.). San Francisco: Freeman, 1980.

Landauer, T. K. Rate of implicit speech. Perception and Psychophysics, 1962, 15, 646.

Massaro, D. W. Preperceptual auditory images. Journal of Experimental Psychology, 1970, 85, 411-417.

Melton, A. Implications of short-term memory for a general theory of memory. Journal of Verbal Learning and Verbal Behavior, 1963, 2,1-21.

Michon, J. A. The making of the present: A tutorial review. In J. Requin (Ed.) Attention and performance VII. Hillsdale, N.J.: Erlbaum, 1978.

Murdock, B. B., Jr. The immediate retention of unrelated words. Journal of Experimental Psychology, 1960, 60, 222-234.

Murdock, B. B.,Jr. Short-term retention of single Paired-associates. Psychological Reports, 1961, 8, 280.

Murdock, B. B., Jr. Recent developments in short-term memory. British Journal of Psychology, 1967, 58, 421-433.

Peterson, L. R., & Peterson, M. J. Short- term retention of individual verbal items. Journal of Experimental Psychology, 1959, 58, 193-198.

Raymond, B. Short-term storage and long-term storage in free recall. Journal of Verbal Learning and Verbal Behavior, 1969, 8, 567-574.

Russo, J. E. Adaptation of cognitive processes to the eye-movement system. In J. W. Senders, D. F. Fisher, & R. A. Monty (Eds.), Eye movements and the higher psychological functions. Hillsdale, N. J.:Eribaum, 1978.

Sperling, G. The information available in brief visual presentations. Psychological. Monographs, 1960, 74 (11, Whole No. 498).

Sperling, G. A model for visual memory tasks. Human Factors, 1963, 5, 19-31.

Sternberg, S. Memory scanning: New findings and current controversies. Quarterly Journal of Experimental Psychology, 1975, 27, 1-32.

Tulving, E., & Colatla, V. Free recall of tringual lists. Cognitive Psychology, 1970, 1, 86-98.

Waugh, N. C., & Norman, D. A. Primary memory. Psychological Review, 1965, 72, 89-104.

Welford, A. T. Fundamentals of skill. London: Methuen, 1968.

Welford, A. T. Attention, strategy and reaction time: A tentative metric. In S. Kornblum (Ed.), Attention and performance. IV. New York: Academic, 1973.

Section 4
Annexe 2

Les principes additionnels d'opération du processeur humain

Il existe des principes additionnels d'opération du processeur humain. Ces principes sont présentés en anglais tels qu'ils sont présentés par les auteurs Card, Moran et Newell [5] dans la publication originale de leurs travaux:

The Model Human Processor- Additional Principles of Operation

P0. Recognize- Act Cycle of the Cognitive Processor.
On each cycle of the cognitive processor, the contents of working memory initiate actions associatively linked to them in long- term memory; these actions in turn modify the contents of working memory.

P1. Variable Perceptual Processor Rate Principle.
The perceptual processor cycle time τ_p varies inversely with stimulus intensity.

P2. Encoding Specificity Principle.
Specific encoding operations performed on what is perceived determine what is stored, and what is stored determines what retrieval cues are effective in providing access to what is stored.

P3. Discrimination Principle.
The difficulty of memory retrieval is determined by the candidates that exist in the memory, relative to the retrieval clues.

P4. Variable Cognitive Processor Rate Principle.
The cognitive processor cycle time τ_C is shorter when greater effort is induced by increased task demands or information loads; it also diminishes with practice.

P5. Fitts's Law.

The Time T_{pos} to move the hand to a target of size S that lies a distance D away is given by

$T_{pos} = I_M \log_2 (DIS + 0.5)$,

where $I_M = 100$ [70~120] msec/bit.

P6. Power Law of Practice.

The time T_n to perform a task on the nth trial folllows a power law

$T_n = T_1 n^{-a}$,

where a = 0.4 [0.2~0.6].

P7. Information Theory Principle.

Decision time T increases with uncertainty about the judgment or decision to be made:

$T = I_C H$,

where H is the information- theoretic entropy of the decision and I_C = 150 [0~157] msec/bit. For n equally probable alternatives (called Hick's law),

$H = \log_2 (n +1)$.

For n alternatives with different probabilities, p_i of occurrence,

$H = \Sigma_i pi \log_2 (1/p_i +1)$.

P8. Rationality Principle.

People act so as to attain their goals through rational action, given the structure of the task and their inputs of information and bounded by limitations on their knowledge and processing abilities:

goals + task + operators + inputs + knowledge + processing limits $\rightarrow$ behavior.

P9. Problem Space Principle.

The rational activity in which people engage to solve a problem can be described in terms of (1) a set of states of Knowledge, (2) operators for changing one state into another, (3) constraints on applying operators, and (4) control knowledge for deciding which operator to apply next.

Section 4
Annexe 3

Exemple de l'usage de la procédure de l'analyse de tâche par le modèle GOMS en utilisant NGOMSL

Dans cette annexe, un extrait de la publication de Keiras [6] est présenté. Le but est de donner un exemple de l'usage de la procédure de l'analyse de tâche par le modèle GOMS en utilisation NGOMSL.

An Example of Using the Procedure

This example shows the use of NGOMSL notation and the top- down breadth-first approach to constructing a GOMS model. The example task and system is text editing with MacWrite, with only one type of task, moving a piece of text from one place to another, analyzed fully. The example consists of four passes over the method description, each pass corresponding to a deeper level of analysis. After the last pass is an illustration of how the methods would be modified to make WM use explicit. Then the operators, task description, and judgment calls are listed, followed by an example sensitivity consideration. For brevity only the new or modified parts of the GOMS model are shown in each pass. The complete example GOMS model can be assembled by collecting the final version of each method.

Description of Methods

Pass1

The topmost user's goal is editing the document. The analysis starts with the unit- task method and the selection rule set that dispatches control to the appropriate method:

Method to accomplish goal of editing the document
Step 1. Get next unit task information from marked-up manuscript.
Step 2. Decide: If no more unit tasks,, then report goal accomplished.
Step 3. Accomplish the goal of moving to the unit task location.
Step 4. Accomplish the goal of performing the unit task.
Step 5. Goto1.

Selection rule set for the goal of performing the unit task.

If the task is moving text, then accomplish the goal of moving text.

If the task is deletion, then accomplish the goal of deleting text.

If the task is copying, then accomplish the goal of copying text.

…etc …

Report goal accomplished.

Pass 2
The recursive procedure begins with the current set of top-level goals being moved to the unit task location and moving text. For brevity, the method for moving the goal to the unit task location, will not be further expanded. This method also provides an example of accessing the task description (the marked-up manuscript).

Method to accomplish the goal of moving to the unit task location
Step 1. Get location of unit task from manuscript.
Step 2. Decide: If unit task location on screen, then report goal accomplished.
Step 3. Use scroll bar to advance text.
Step 4. Goto2.

The method for moving text involves a judgement call of assuming that users view moving text as first cutting, then pasting:

Method to accomplish goal of moving text
Step 1. Cut text
Step 2. Paste text
Step 3. Verify correct text moved.
Step 4. Report goal accomplished.

Step 1 and Step 2are represented here temporarily with high- level operators, which in the next pass, will be replaced with **Accomplish Goal** operators, and methods provided. Notice that Step 3 assumes that the user will pause to verify that the desired results have been obtained. This analysis assumes (perhaps wrongly) that a similar verification is not done within the cutting and pasting methods to be described below.

Pass3

this pass provides the methods for cutting and pasting. Notice below how Steps 2 and 3 of the moving text method have been changed from the previous pass. To illustrate the guidelines, the first draft of the method for cutting is too long; in response to the guideline advice, this is fixed in the second draft.

Method to accomplish goal of moving text
Step 1. Accomplish the of cutting text
Step 2. Accomplish the goal of pasting text
Step 3. Verify correct text moved.
Step 4. Report goal accomplished.

Method to accomplish goal of cutting text- First Draft
Step 1. Move cursor to beginning of text.
Step 2. Hold down mouse button.
Step 3. Move cursor to end of text.
Step 4. Release mouse button.
Step 5. Move cursor to EDIT menu bar item.
Step 6. Hold down mouse button.
Step 7. Move cursor to CUT item.
Step 8. Release cursor button.
Step 9. Report goal accomplished.

Notice that this last method is correctly described, but it has too many steps. Also, this only the second level of goals, and the method already has motor primitives. Notice that Steps 1-4 correspond to a general method for how things are selected almost everywhere on the Macintosh, and Steps 5-8 are involved with issuing the CUT command. Perhaps the analysis has stumbled into providing a trace- based method for executing a specific task rather than general methods that cover the tasks of interest, as discussed above. The second draft of the method corrects the problems with the judgment calls the users know and take advantage of the general selecting function, and so they will have a "subroutine" method for selecting text, and that similarly, they also have a general method for issuing commands. These two sequences in the first draft can be collapsed into two high- level operators, as shown in second draft below of the cutting method. The pasting method is then written in a similar way.

Method to accomplish goal of cutting text- Second Draft
Step 1. Select text.
Step 2. Issue CUT command.
Step 3. Report goal accomplished.

Method to accomplish goal of pasting text
Step 1. Select insertion point.

Step 2. Issue PASTE command.
Step 3. Report goal accomplished.

Pass4
This pass provides methods for selecting text and the corresponding selection rules, and also methods for selecting the insertion point and issuing the CUT and PASTE commands. A simplifying assumption is made that the user does not make use of the command- key shortcut.

Method to accomplish goal of cutting text
Step 1. Accomplish goal of selecting text.
Step 2. Accomplish goal of issuing CUT command.
Step 3. Report goal accomplished.

Method to accomplish goal of pasting text
Step 1. Accomplish goal of selecting insertion point.
Step 2. Accomplish goal of issuing PASTE command.
Step 3. Report goal accomplished.

Selection rule set for goal of selecting text

If text- is word, then accomplish goal of selecting word.

If text- is arbitrary, then accomplish goal of selecting arbitrary text.

Report goal accomplished.

Method to accomplish goal of selecting word
Step 1. Determine position of beginning of word.
Step 2. Move cursor to beginning of word.
Step 3. Double- click mouse button
Step 4. Verify that correct text is selected
Step 5. Report goal accomplished.
Method to accomplish goal of selection arbitrary text

Step 1. Determine position of beginning of text.
Step 2. Move cursor to beginning of text.
Step 3. Press mouse button down.
Step 4. Determine position of end of text .
Step 5. Move cursor to end of text.
Step 6. Verify that correct text is selected.
Step 7. Release mouse button.
Step 8. Report goal accomplished.

The last method above seems to be too long. This is the result of a judgment call that the has to look at the marked-up manuscript to see where the text starts and then find this spot on the screen (Step1), and then as a separate unit of activity, and move the cursor there (Step 2). A similar situation appears in Steps 4 and 5. Some alternative judgment calls: Perhaps there is a "drag" operator and using it requires determining the end of the text before pressing down the mouse button. Alternately, perhaps the sequence appearing in Steps 1 and 2 and Steps 4 and 5 corresponds to a natural goal of "find a place and put the cursor there," for which there should be a high- level operator and later a method. For brevity, these alternative judgment calls are not pursued in this example. The remaining methods are as follows:

Table 1: Analyst- Defined Operators for the Example

Get next unit task information from marked-up manuscript- look at manuscript and scan for the next edit marking, and put some of the task description into working memory.

No more unit tasks- tells whether there was another edit marking.

Task is … - tells whether task is of the specified type, such as move, copy, etc.

Get location of unit task from manuscript- look at edit marking on manuscript and determine position in the manuscript.

If unit task location on screen- tells whether the material corresponding to the edit marking is on the screen.

Use scroll bar to advance text- a high- level operator that could be expanded into a set of methods.

Determine position of- get information from task description, and map to perceptual location on screen.

Text-is- tells whether text is a word, sentence, or arbitrary.

Verify- compare results to goal to check that desired result is achieved.

Move cursor to- move mouse until cursor at specified point.

Click mouse button.

Double-click mouse button.

Press mouse button.

Release mouse button.

Method to accomplish goal of selecting insertion point
Step 1. Determine position of insertion point.
Step 2. Move cursor to insertion point.
Step 3. Click mouse button.
Step 4. Report goal accomplished.

Method to accomplish goal of issuing CUT command (assuming that user does not use command-X shortcut)
Step 1. Move cursor to "Edit" on Menu Bar
Step 2. Press mouse button down.
Step 3. Move cursor to "CUT".
Step 4. Verify that CUT is selected.
Step 5. Release mouse button.
Step 6. Report goal accomplished.

Method to accomplish goal of issuing PASTE command (assuming that user does not use command-V shortcut)
Step 1. Move cursor to "Edit" on Menu Bar
Step 2. Press mouse button down.
Step 3. Move cursor to "PASTE".
Step 4. Verify that PASTE is selected.
Step 5. Release mouse button.
Step 6. Report goal accomplished.

Modifications to Show WM Usage
Some of the methods from the last pass of the above example will be written to illustrate how WM usage is made explicit. The result is a generic sub-method for issuing a command that captures some of the consistency of the Macintosh menu- based command interface.
The first illustration will use the informal approach. Instead of separate methods for issuing a CUT and a PASTE command whose name has been deposited previously in WM. First, the methods that called the previous command- issuing methods need to be modified to put the command name in WM, and to accomplish the generic goal of issuing a command.

Method to accomplish goal of cutting text
Step 1. Accomplish goal of Selecting text.

Step 2. Retain that command is CUT, and accomplish goal of issuing a command.
Step 3. Report goal accomlish goal accomplished.

Method to accomplish goal of pasting text
Step1. Accomplish goal of Selection point.
Step2. Retain that the command is PASTE, and accomplish goal of issuing a command.
Step 3. Report goal accomplished.

The following generic command-issuing method assumes that the user must remember with a retrieval from LTM which one of the menus contains the command, and that the user has to remember this menu name in WM while executing the method.

Method to accomplish goal of issuing a command
Step 1. Recall command name, and retrieve from LTM the menu for it, and retain the menu name.
Step 2. Recall the menu name, and move cursor to it on Menu Bar.
Step 3. Press mouse button down.
Step 4. Recall command name, and move cursor to it.
Step 4. Recall command name, and verify that it is selected.
Step 5. Release mouse button.
Step 6. Forget menu name, forget command name, and report goal accomplished.

Table 2: Example Task Description
Task is to move specified piece of text
Price of text is a word, or arbitrary
Position of beginning of text
Position of end of text, if it is arbitrary
Position of destination
The following is the same generic command- issuing method, but using the formal variable approach instead.

Method to accomplish goal of issuing a command
Step 1. Recall that command name is X retrieve from LTM that Y is the menu name for X and retain that menu name is Y.

Step 2. Recall that menu name is X, and move cursor to X on Menu Bar.
Step 3. Press mouse button down.
Step 4. Recall that command name is X, and move cursor to X.

Step 4. Recall that command name is X, and verify that X is selected.
Step 5. Release mouse button.
Step 6. Forget menu name, forget command name, and report
goal accomplished.

Competing the Analysis

Table 1 shows the list of analyst- defined operators, and Table 2 shows the information for this GOMS model. The assumptions and judgment calls made are listed in Table 3.

Checking Sensitivity to Judgment Calls

as an example discussion, suppose the user does not decompose the move task into cut-then-paste as was assumed, but thinks of move as a single goal. An alternative design would thus have the user selecting the the text, issuing a move command, and then clicking the mouse where the text is to be moved. The methods tailored to this alternative judgment of how move and copy goals decompose would probably show is that more methods might be needed overall because the cut and paste sub-methods could not be shared with other editing methods. So the quality of the design in terms of its learnability and consistency is probably sensitive to whether the judgment call is correct. The analyst may want to explore these alternatives by comparing the two designs in detail.

Section 4
Annexe 4

Bibliographie de la section 4

[1] Anderson, J.R. (1982). Acquisition of cognitive skill. Psychological review, 89,369-406.

[2] Bennett, J.L., Lorch, D.J., Kieras, D.E., & Polson, P.G. (1987). Developing a user interface technology for use in industry. In Bullinger, H.J., & shackel, B. (Eds.), Human_ Computer Interaction_ INTERACT, 87. Amsterdam: North_Holland.

[3] Bjork, R.A.(1972). Theoretical implications of directed forgetting. In A.W. Melton and E. Martin (Eds.), Coding Processes in Human Memory. Washington, D.C.: Winston.

[4] Bovair, S., Kieras, D.E., & Polson, P.G. (1988). The acquision and Performance of text_editing skill: A Production_ system analysis. (Technical Report No.28), University of Michigan, Ann Arbor.

[5] Card, S., Moran, T. & Newell, A. (1983). The Psychology of Human-Computer Interaction. Hillsdale, New Jersey: Eribaum.

[6] Keiras D.E. (1988). Towards a practical GOMS model methodology for user interface design. In M. Helander (Eds.), Handbook of Human-Computer Interaction, p135-157. New York: North-Holland

[7] Kieras, D.E. (1986). A mental model in user-device interaction: A Production system analysis of a problem solving task. Unpublished manuscript, University of Michigan.

[8] Kieras, D.E. (1988). Making cognitive complexity practical. In CHI,88 Workshop on Analytical Models, Washington, May 15,1988.

[9] Kieras, D.E. (in press). The role of cognitive simulation models in the development of advanced training and testing systems. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), Diagnostic Monitoring of Skill and Knowledge Acquisition. Hillsdale, N.J.:Erlbaum.

[10] Kieras, D.E., & Bovair, S. (1986). The acquisition of procedures from text: A production system analysis of transfer of training. Journal of Memory In Language, 25,507_524.

[11] Kieras, D.E. & Polson, P.G. (1985). An approach to the formal analysis of user complexity. International Journal of Man_Machine Studies,22,365_394.

[12] Polson, P.G. (1987). A quantitative model of humancomputer interction. In J.M. Carroll (Ed.): Interfacing Thought: Cognitive Aspects of Human-Computer Interaction. Cambridge, MA: Bradford, MTT press.

Section 5

Conclusion et perspectives

Chapitre 11 : Conclusion et perspectives

Chapitre 11

Conclusion et perspectives

L'introduction de la parole dans les interfaces homme - machine ouvre des perspectives intéressantes pour la communication multimodale malgré les performances encore limitées de la reconnaissance automatique de la parole et la position forte de l'utilisation des modalités gestuelles habituelles au niveau d'entrée.

L'étude de la parole que nous avons faite a porté sur l'intégration de ce mode d'interaction dans les interfaces existantes. Globalement, les interfaces multimodales semblent être plus intéressantes que les interfaces exclusivement graphiques.

Par rapport à des résultats d'autres auteurs, [10] par exemple, l'usage de la multimodalité "synergique" semble plus important. Nous avons observé une spécialisation des modes selon la tâche ainsi qu'une tendance à fiabiliser l'interaction par l'usage de la redondance. Ce facteur prend moins d'importance si l'action est complexe ou à faible risque : il semble que des considérations d'économie l'emportent alors pour limiter l'usage de la redondance.

A ce stade de notre investigation nous avons pu juger positivement de l'intérêt des propriétés CARE pour situer l'usager par rapport à un contexte d'interaction.

Nous pouvons détailler notre apport sur trois axes principaux :

Axe 1 : L'usage de la multimodalité. La première question qui se posait était : qu'est-ce qu'on peut faire avec une interface multimodale et comment l'utilisateur réagit- il quand il se trouve en situation d'interaction multimodale.

Cette question a été le sujet des chapitres 4 et 5. Et pour repondre à cette question, deux experiences ont été réalisées. Dans le premier corpus, l'usager est en situation de communication entre humains. L'usager a le libre choix entre soit utiliser un mode gestuel de type *faire*, soit de s'orienter vers l'assistant humain pour lui demander, donc avec la parole en mode *faire-faire* - en disposant d'une syntaxe limitée des mots – d'accomplir la tâche.

Les résultats de cette expérience ont montré que la multimodalité c'est à-dire ses deux modes sont autant utilisés par l'usager.

Dans le deuxième corpus, l'usager se trouve en communication avec la machine sur laquelle tourne un logiciel multimodal. Malgré les contraintes techniques et la performance moyenne des logiciels utilisés, la multimodalité a aussi été largement utilisée, même pour le lot de sujets considéré, des experts que l'on pouvait considérer comme attachés à leurs habitudes anciennes.

Dans les deux cas, la multimodalité est acceptable dans l'interaction homme – machine. Ces expériences étaient faites dans un environnement de « laboratoire » et avaient le but de montrer les capacités offertes par l'intégration de la parole comme mode d'interaction.

Axe 2 : L'utilisabilité de la multimodalité. La deuxième question qui se pose est : comment peut-on utiliser la multimodalité dans les applications réelles ? Cette question a été le sujet d'une expérience qui était faite dans un environnement réel de travail (chapitre 7) avec un logiciel de reconnaissance commercial intégré dans une application d'édition de texte. Les usagers sont des experts qui exercent leur travail ordinaire sur cette application.

Même en considérant les effets d'une utilisation "par curiosité", l'expérience a montré l'intérêt de la multimodalité pour une bonne partie des sujets,.

Axe 3 : Le coût de la multimodalité. La troisième question porte sur le coût (cognitif et moteur) de la multimodalité. Cette question a été le sujet d'une étude expérimentale temporelle de la multimodalité, faite sur des sujets experts bien entraînés à l'usage de la multimodalité. L'expérience avait pour but de mesurer les effets de la multimodalité sur le temps nécessaire pour accomplir certaines tâches.

La multimodalité diminue pour beaucoup de sujets, le temps global de la tâche. Mais elle a un coût de traitement du parallélisme qui peut s'ajouter, chez certains sujets, au temps global de l'exécution et qui fait que le profit de l'interaction multimodale n'est pas maximal.

Nous pouvons donc faire les recommandations suivantes au concepteur de systèmes multimodaux :

- L'introduction de la multimodalité est souhaitable pour tenir compte de la diversité des usages et des préférences des utilisateurs. Ceci est d'autant plus vrai pour les systèmes à large diffusion.

- Une grande attention doit être apportée à l'adéquation de la modalité orale du côté du vocabulaire et du côté de la consistance car l'usager a tendance à abandonner rapidement la modalité orale lorsqu'il rencontre des difficultés pour la mettre en œuvre ou lorsqu'il se trouve en situation d'échec.
- L'utilisabilité des interfaces multimodales reste conditionnée par les performances des techniques utilisées dans les composants de l'interface, notamment pour la parole. Il y a donc lieu de tester ces performances sur le lieu et dans les conditions réelles de travail.

Enfin on peut penser que la parole étant un mode de communication naturelle, a les atouts pour s'imposer comme un mode de communication avec la machine. Nos études confirment ce point de vue.

RESUME
Le but de ce travail est d'étudier l'introduction de la parole dans les interactions homme-machine multimodales et de trouver des contraintes de spécification d'une interface multimodale qui rende la communication avec la machine plus efficace et moins coûteuse. Les objectifs de ce travail concernent les trois points suivants : l'étude de l'usage de la multimodalité, la réalisation d'un système multimodal et l'étude de l'utilisabilité de la multimodalité.
L'étude a utilisé des méthodes expérimentales pour clarifier ces points. En ce qui concerne l'usage de la multimodalité deux expériences sont effectuées sur des sujets de test : d'abord dans un cadre de communication inter-humaine qui simule une interface homme-machine réelle, puis en situation de communication homme-machine réelle avec une interface multimodale. Le but de ces expériences est d'observer ce que l'interface multimodale peut offrir et quelles sont ses limites.
En ce qui concerne l'utilisabilité de la multimodalité, des tests sont faits à partir d'une expérimentation d'une interface homme - machine multimodale. L'étude se base sur un modèle cognitif de l'être humain pour définir et évaluer les interfaces homme-machine.

MOTS-CLES
communication homme machine, interaction homme-machine, interface multimodale, parole, GOMS, NGOMSL.

Speech Usage In Multimodal Interfaces

Abstract
The goal of this thesis is to consider the issues related to speech introduction in multimodal human - machine interfaces. To determine and establish the problems and constraints of specification, which can make human-machine communication to maximum efficiency. The study is concentrated on the user to find what impacts exist when using multiple communication channels in human-machine interaction.
Main issues covered the usage of multimodality, the implementation of multimodal human-machine interfaces and the usability of multimodality.
Experimental methods were tested on 34 users, who performed two simulations. First, in inter-human interaction, as a simulation of the human-computer interface. Second, in real human-computer communication environment, using a multimodal interface. These experiments where aimed at studying what we can be achieved by using a multimodal interface and to determine its limitations.
A second experimental study will be aimed at the usability of a multimodal interface in real live applications. Studying the behavior of the users in relation to the facilities used, the environment, and the performance of the real multimodal interface in this situation.
Finally, using the experimental methods and a cognitive model of the human user, this thesis is aimed to prove that the use of multimodality involves a certain time overhead, due to the effects of the concurrent use of human communication channels.

Keywords
computer human interface, multimodality, multimodal interface, speech, GOMS, NGOMSL.

Laboratoire CLIPS (Communication Langagière et Interaction Personne – Système) Institut IMAG – Domaine Universitaire BP53, 38041 Grenoble cedex - FRANCE