

Les déterminants du choix d'une modalité d'interaction avec une interface multimodale

Laurent Karsenty

INTUILAB, Prologue 1, La Pyrénéenne, 31672 Labège
karsenty@intuilab.com

RESUME

Les applications industrielles des interfaces multimodales touchent de plus en plus de domaines. Pourtant, leur conception soulève encore de nombreuses questions, en partie liées à notre méconnaissance des déterminants du choix d'une modalité d'interaction par l'utilisateur. Cet article rapporte une série d'observations tirées d'expérimentations d'interfaces multimodales qui comblent en partie ce manque. L'effet de certains déterminants est spécifiquement discuté, comme la nature de la tâche et l'évaluation subjective et contextualisée de la performance réalisée par l'utilisateur alors qu'il utilise le système.

MOTS CLES : multimodalité, commande vocale, geste

ABSTRACT

Multimodal interfaces get wider recognition in industry and more and more applications are considered today. However, their design still raises many questions, partly related to our lack of knowledge regarding the determinants of the choice of an interaction modality by users. This article reports a series of observations on the use of multimodal interfaces which partly fill this lack. The effect of some determinants is specifically discussed, such as the nature of the task and the subjective and contextualized evaluation of the performance carried out by users as they are using the system.

KEYWORDS : multimodality, spoken command, gesture

INTRODUCTION

Les développements récents de l'informatique et des infrastructures de télécommunication ouvrent aujourd'hui la porte à une interaction plus riche entre l'homme et la machine. En particulier, il devient possible d'avoir recours à une interaction multimodale, dépassant le paradigme clavier-souris-écran pour exploiter plusieurs modalités d'interaction, que ce soit sur un poste fixe ou un terminal mobile. Ces perspectives concernent des applications aussi variées que les bornes publiques d'information (ex., Lamel et al., 2002), les services mobiles de recherche d'information (Karsenty et al., 2005), les systèmes embarqués (ex., Kamp, 1998), les environnements

collaboratifs et les applications de réalité virtuelle (Thalman, 2002), pour ne citer que quelques exemples.

Si les premiers travaux significatifs sur la multimodalité ne datent pas d'aujourd'hui, force est de reconnaître que la conception d'interfaces multimodales soulève encore de nombreuses questions. Pour une part, ces questions sont liées à la méconnaissance des déterminants du choix de chaque modalité d'interaction ou combinaison de modalités par l'utilisateur. Pourtant, en analysant les travaux empiriques à notre disposition et en recoupant plusieurs résultats disparates, on s'aperçoit qu'une certaine connaissance est déjà disponible laquelle, si elle est encore insuffisante pour établir une théorie complète et précisément guider les choix de conception, peut néanmoins faciliter certaines décisions des concepteurs. C'est l'objectif de cet article que de rapporter les résultats d'une telle analyse sous forme d'une synthèse des principales observations recueillies sur l'usage d'interfaces multimodales. Cette synthèse permettra de tirer une série d'enseignements et de pistes de réflexion. Pour respecter les limites imposées à cet article, nous nous focaliserons dans la suite sur la multimodalité *en entrée*. L'usage des modalités vocale et gestuelle sera plus particulièrement traité car elles sont au cœur de nombreuses applications envisagées aujourd'hui. Par gestes, nous entendrons à la fois les gestes 2D ou 3D mais aussi les gestes de manipulation d'une souris, d'un stylet ou de tout autre dispositif de commande.

QUELQUES DEFINITIONS

Système multimodal

Nous considérerons comme système multimodal tout permettant d'utiliser plusieurs canaux de communication en entrée et/ou en sortie (Nigay et Coutaz, 1996, Martin, Julia et Cheyer, 1998). Nous retiendrons en outre que les systèmes multimodaux offrent une autre source de flexibilité aux utilisateurs : elle concerne *le format* des informations transmises ou, ce que Bernsen (2002) a proposé d'appeler la *modalité représentationnelle*. Pour chaque canal, il existe en effet plusieurs façons de « conditionner » l'information à transmettre. Par exemple, en s'adressant à un serveur vocal, l'utilisateur peut formater son message en une suite de mots-clés, de morceaux de phrase simplifiée ou en une expression totalement naturelle. Nous verrons que du point de vue de l'utilisateur, ces deux dimensions, canal de

communication et modalité représentationnelle, ne sont pas indépendantes.

Modes de coordination entre modalités

La transmission d'une information peut s'appuyer sur une modalité unique ou sur une combinaison de modalités. Ainsi, pour indiquer une destination souhaitée, on peut dire « Je voudrais aller au 5 rue du Pont » ou désigner le lieu souhaité tout en disant « Je voudrais aller *ici*. » Avec des systèmes où une combinaison de modalités en entrée est possible, l'utilisateur doit connaître *les modes de coordination entre modalités* autorisés (Coutaz & Nigay, 1994).

L'usage de plusieurs modalités peut être *séquentiel* ou *parallèle*. Certains systèmes n'autorisent que des combinaisons séquentielles de modalités, sans recouvrement dans le temps. C'est le cas notamment du prototype MiPad (Huang et al., 2001) où l'utilisateur, pour remplir un champ d'information, doit d'abord cliquer dessus puis parler. D'autres systèmes autorisent des recouvrements de modalités, en faisant intervenir un mécanisme explicite ou implicite de fusion multimodale combinant le sens extrait de chaque modalité. Par exemple, avec le système Intuikit développé par Intuilab¹, l'utilisateur peut saisir un objet représentant une valeur boursière avec sa main sur un écran tactile et dire en même temps « Je voudrais en acheter pour 200 euros », par exemple.

Quel que soit le mode disponible ou choisi, les modalités peuvent être utilisées de façon *complémentaire*, *redondante* ou même *concurrente*. Deux modalités sont complémentaires si les informations que chacune véhicule constituent une partie du sens global communiqué. Deux modalités utilisées de façon séquentielle ou parallèle sont redondantes si les informations que chacune véhicule sont identiques. Par exemple, un utilisateur peut dire : « Je veux aller à Toulouse » tout en montrant la ville de Toulouse sur une carte. Enfin, deux modalités peuvent être utilisées de façon concurrente, au sens de contradictoire ou conflictuelle. Par exemple, un utilisateur pourrait lancer la commande vocale « tourne à droite » en désignant gestuellement la gauche.

Enfin, deux modalités sont *équivalentes* si l'utilisateur ou le système les utilise indifféremment pour transmettre la même information. Dans le cas inverse, on peut parler de *l'assignation* d'un type d'information à une modalité donnée.

Toutes ces distinctions sont fondamentales pour l'étude des usages de la multimodalité. Si un système offre plusieurs modalités d'entrée, on peut en effet se demander s'il y a intérêt à fournir des mécanismes de fusion multimodale supposant que l'utilisateur agisse en combinant deux modalités (à peu près) en même temps ou si celui-ci ne trouverait pas plus facile et/ou plus

efficace de les enchaîner séquentiellement. On peut aussi se demander s'il préférerait tirer parti des spécificités de chaque modalité et les combiner de façon complémentaire ou, au contraire, chercher à réaliser un acte complet avec seulement une modalité, quitte à être parfois moins efficace. Il est aussi nécessaire de savoir si l'utilisateur adopterait plusieurs modalités indifféremment pour transmettre une même information ou si, au contraire, il ne préférerait pas toujours la même modalité pour une fonctionnalité ou une commande donnée, auquel cas il serait inutile de fournir plusieurs modalités. Enfin, il convient de déterminer s'il n'existe pas plusieurs profils d'utilisateurs face à ces nouvelles interfaces et ce qui les caractériserait.

USAGES OBSERVÉS LORS D'EXPERIMENTATIONS

Une synthèse des résultats de plusieurs expérimentations est présentée ici. Soulignons que beaucoup de ces travaux de recherche portent sur des environnements simulants en partie les capacités du système. Pour cette raison, une attitude prudente dans l'interprétation des résultats reste nécessaire.

Commandes multimodales vs. monomodales

Plusieurs travaux comparent le taux de commandes multimodales au taux de commandes monomodales pour vérifier l'intérêt des premières. Généralement, le terme de commande multimodale y désigne une utilisation complémentaire et parallèle des modalités. De façon générale, on constate *la prédominance de commandes uniquement vocales* : ainsi, dans des tâches exploitant le plan d'une ville, Oviatt et al. (1997) enregistre 63,5% de commandes effectuées oralement, contre 17,5% effectuées par l'écriture et 19% par des commandes multimodales associant parole et stylet. Dans une tâche de manipulation de fichiers, la même prédominance a été observée (Huls & Bos, 1998) : environ 58% des commandes étaient réalisées à la voix uniquement, alors que seulement 11% environ étaient réalisées de façon multimodale.

Toutefois, ces résultats varient en fonction des *tâches* réalisées par les utilisateurs. Ainsi, dans une tâche d'aménagement d'un salon, impliquant la manipulation et l'organisation d'objets dans l'espace (canapé, table, chaises, etc.), Mignot et Carbonel (1996) ont enregistré 43% de commandes vocales, contre 41% de commandes multimodales (geste+voix) et 16% de commandes uniquement gestuelles. Oviatt et al. (1997) ont constaté de leur côté que les commandes multimodales étaient produites plus fréquemment pour des tâches spatiales de localisation (ajouter, déplacer, modifier ou calculer la distance entre les objets). Enfin, Catinis et Caelen (1995) enregistrent, sur une tâche de dessin, 66% de commandes multimodales. Cela dit, dans ce dernier cas, l'utilisation de la commande vocale était très contrainte : une seule structure syntaxique et un vocabulaire de quelques dizaines de mots seulement étaient autorisés. Cette spécificité, qui a pu favoriser le recours à la

¹ On pourra consulter une démo de cette application à : <http://www.intuilab.com/presentation/202-demos.html>

multimodalité, révèle que l'adoption d'une modalité au sens de canal de communication peut dépendre à la fois de la tâche mais aussi du format d'information autorisé.

De plus, certaines conditions externes favoriseraient le recours à l'une ou l'autre des modalités. Huls et Bos (1998) ont notamment révélé l'effet de deux facteurs sur ce choix dans une tâche de manipulation de fichiers informatiques (des résultats similaires sont présentés dans Oviatt et al., 1997) :

- *La longueur des mots* (noms de fichier) : les mots courts – 3 à 4 caractères dans cette étude – favorisaient le recours à la modalité vocale (76% des commandes) alors que les mots longs – plus de 8 caractères dans cette même étude – conduisaient à un ré-équilibre dans le choix des commandes avec 39% de commandes de type manipulation directe (avec la souris), 18% de commandes multimodales (parole + désignation avec la souris) et 43% de commandes vocales.
- *L'accessibilité visuelle des objets* : dans l'expérimentation de Huls et Bos, les fichiers à manipuler pouvaient être visibles à l'écran ou non. Lorsqu'ils n'étaient pas visibles, les utilisateurs avaient préférentiellement recours à la modalité vocale seule pour y faire référence (72% des cas). Par contre, lorsqu'ils étaient visibles, la modalité vocale seule ne représentait plus que 44% des cas.

Bien évidemment, la combinaison de ces deux facteurs a produit des résultats encore plus marqués puisque, dans le cas de noms courts et non visibles, 86% des commandes utilisaient la parole seule, alors qu'avec des noms longs visibles, la commande vocale ne représentait plus que 28% des cas.

D'autres conditions favoriseraient l'adoption d'une modalité en particulier au détriment des autres. C'est le cas notamment de la modalité de sortie utilisée par le système. Ainsi, on a observé qu'un prompt vocal favorisait le choix de la modalité vocale en entrée alors que plusieurs modalités étaient disponibles (Lamel et al., 2002). On peut supposer qu'une forme de pré-activation intra-modale est possible entre la modalité utilisée en entrée et celle utilisée en sortie.

Modifications du langage dans un environnement multimodal

Plusieurs auteurs ont montré que l'utilisation de plusieurs modalités modifiait le comportement langagier. Les modifications suivantes ont notamment été relevées :

- *Simplification du langage* : les expressions produites dans un environnement multimodal sont plus brèves et moins complexes que les énoncés unimodaux (Oviatt et al., 1997, Petrelli et al., 1997). Par exemple, si avec la modalité vocale seule, l'utilisateur dit (Oviatt et al., 1997) : « Je voudrais voir la photo de la maison à l'extrémité sud-ouest de Reward Lake », dans un environnement multimodal, il encerclerait la maison avec son stylet tout en disant : "Montre la photo". Les études citées ne

disent pas, par contre, si une telle simplification est observée chez tous les utilisateurs. Par contre, ils mentionnent qu'elle implique une amélioration des taux de reconnaissance correcte des commandes vocales.

- *Antériorité du composant de localisation avec le geste* : dans les énoncés vocaux, le constituant de localisation est rarement au début (seulement dans 1% des énoncés) ; par contre, les expressions multimodales débutent invariablement avec un geste graphique délivrant l'information de localisation suivi d'un énoncé verbal (Oviatt et al., 1997) . Par exemple, l'utilisateur dessine un cercle et dit : « Ajoute une piscine », alors qu'avec la voix seule, il dirait plus probablement : « Ajoute une piscine + [énoncé de localisation] ». Il faut souligner que l'antériorité du geste sur la parole a été observée systématiquement dans les études sur la communication humaine (cf. Butterworth & Hadar, 1989) et devrait reposer sur des dispositions naturelles mises en œuvre en planifiant un acte multimodal.
- *Apparition de nouvelles formes de référence*. Plusieurs études font mention notamment des formes suivantes (Siroux et al., 1995, Petrelli et al., 1997) :
 - des références déictiques combinées à un geste désignant une localisation (ex. « Y a-t-il des campings ici ? ») ;
 - l'utilisation redondante de références définies vocales avec un geste déictique (ex. « Donne moi la distance entre Lansing et Morestel » avec une désignation tactile sur chacune des deux villes) ;
 - des références mixtes, avec une désignation tactile complétant une requête verbale sans référence à un lieu (ex. « Quels sont les campings » + une désignation sur une localisation). Toutefois, il convient de noter que ce dernier type de combinaison est soit observé rarement, soit observé avec des utilisateurs expérimentés. Il ne semble donc pas « naturel » mais pourrait apparaître intéressant avec l'expérience.
- *Conception des expressions référentielles en fonction des caractéristiques perceptives* : l'exploitation d'un support visuel peut aussi modifier le langage et, en particulier, les expressions référentielles (Gaiffe & Romary, 1997). Ainsi, l'absence de structuration spatiale et/ou de caractéristiques perceptives contrastant des objets représentés induit des expressions complexes et variables d'un utilisateur à l'autre. Par exemple, pour désigner un point non identifié sur la carte d'une ville, l'utilisateur peut dire : « ... entre le Pont Charlemagne et l'avenue Gabriel Péri ». Par contre, si la carte est quadrillée, l'utilisateur pourra

employer une expression plus simple telle que «... en A 8». Toutefois, notons que cette analyse ne prévoit pas la difficulté potentielle de l'utilisateur pour savoir quelle(s) caractéristique(s) perceptive(s) il peut exploiter vocalement pour faire référence à un objet présenté visuellement.

Toutes ces observations confirment l'idée que la multimodalité a un réel potentiel pour améliorer la performance des utilisateurs et qu'une partie de ce gain au moins peut être acquis naturellement en vertu du fait que l'action humaine – langagière, manuelle ou autre – est fortement contextualisée.

Usage complémentaire des modalités

La complémentarité a été observée dans plusieurs études. Cette complémentarité concerne le plus souvent la combinaison entre une expression déictique (ici, là, ce, ...) et un geste déictique et sert ainsi à faciliter la localisation. Mais elle peut aussi servir d'autres fonctions. Elle permet notamment de combiner un geste anaphorique ou un geste de manipulation avec de la parole.

Ainsi, avec le système QuickSet (Cohen et al., 1997), l'utilisateur peut contrôler un environnement multimodal de simulation et lancer des commandes du type : « Jeep 23, suivez cette route d'évacuation » avec un geste (stylet) montrant la direction à suivre. Le geste peut ici simplement consister à indiquer par une flèche la direction à suivre : il possède pour cette raison des vertus anaphoriques. Avec le même système, l'utilisateur peut dessiner un carré tout en disant « Crée une unité médicale ici », la parole donnant alors sens au geste de manipulation de l'outil dessin. Avec un utilisateur entraîné à ces styles d'interaction, le temps de création d'entités a été divisé par 9 par rapport à l'utilisation d'une interface graphique standard (Cohen et al., 1998).

Trois types de complémentarité se dessinent donc :

- geste déictique + voix ;
- geste anaphorique + voix ;
- geste de manipulation + voix.

Beaucoup de travaux insistent sur le premier type de complémentarité. Pourtant, il se pourrait bien qu'il n'ait qu'un intérêt relatif dans le cadre de l'interaction homme-machine. Une étude sur l'utilisation d'une carte de ville multimodale dans laquelle les utilisateurs pouvaient créer, déplacer, supprimer des objets ou calculer des distances entre des points (Oviatt et al., 1997), a montré que les expressions multimodales du type geste déictique + voix représentaient 14% de l'ensemble des expressions multimodales, les 86% restant étant constitué d'expressions du type geste de manipulation + voix. Ces résultats semblent indiquer que les utilisateurs exploitent avantagement la multimodalité pour des actions complexes, là où le gain en efficacité devrait être le plus marqué.

Effets de l'expérience

Un fait à mettre en parallèle avec cette apparente recherche d'efficacité est *l'accroissement du taux de commandes multimodales avec l'expérience*. Par exemple, Petrelli et al. (1997) notent que des utilisateurs expérimentés réalisent 84% de leurs commandes de façon multimodale alors qu'ils ne sont que 30% parmi les novices. Des résultats moins impressionnants mais démontrant la même tendance ont été enregistrés par Mignot & Carbonell (1996). Toutefois, cette évolution ne concernait alors qu'une moitié des utilisateurs.

Globalement, ces observations suggèrent que la réalisation de commandes multimodales apparaît soit peu intuitive à l'utilisateur novice, soit difficile à exécuter ne serait-ce que parce qu'elle demande à combiner deux modalités. Ce n'est qu'avec l'expérience, et donc une certaine maîtrise des commandes à modalité unique, que certains d'entre eux se lanceraient à expérimenter les commandes multimodales. Ils auraient alors la possibilité de constater les gains d'efficacité que leur apporte cette nouvelle forme d'interaction.

Usage parallèle ou séquentiel de modalités

Théoriquement, les modalités utilisées de façon complémentaire peuvent être mises en œuvre parallèlement ou séquentiellement. Une question importante est de savoir si, spontanément, l'utilisateur aurait le plus souvent recours à un usage parallèle ou séquentiel des modalités. Cette question a souvent été traitée en analysant l'utilisation conjointe d'un geste déictique avec une référence déictique. En fait, les résultats obtenus à ce jour sont contradictoires :

- Une étude empirique de la temporalité des modalités (Oviatt et al., 1997) a montré que l'usage parallèle des modalités ne constituait pas la majorité des cas. Le parallélisme entre une expression déictique et un geste déictique (« mets ce triangle ici ») ne concernait alors que 25% de ce type de commande. Le plus souvent, le geste précédait la parole, avec un intervalle de temps moyen de 1 seconde, et un espace de variation allant de 0 à 4 secondes maximum.
- D'un autre côté, dans l'expérimentation de Hauptmann & McAvinney (1993) sur des manipulations multimodales de cubes, les gestes étaient réalisés pendant ou après le début de la voix dans 50% des cas. Ils ne commençaient avant la commande vocale que dans 8-9% des cas. Le même type de résultat a été enregistré par Catinis (1998) dans une tâche de dessin.

A côté de cela, il est intéressant de constater que des mesures très proches de celles fournies par Oviatt et al. ont été obtenues à partir d'études de la multimodalité geste + parole dans la communication homme-homme (cf. Butterworth & Hadar, 1989). On pourrait donc penser que la précedence du geste sur la parole a un caractère naturel ou spontané intéressant à reproduire en interaction homme-machine. Mais il est aussi possible

que suivant le dispositif offert aux utilisateurs (type de microphone, type de dispositif de désignation : main ou stylet, type de geste autorisé : 2D ou 3D, etc.) et suivant les tâches étudiées, cette règle ne s'applique pas de façon identique.

Usage redondant des modalités

De manière générale, peu de redondance dans les expressions multimodales a été observée. Toutefois, les observations recueillies ne sont pas toutes homogènes. Ainsi, Petrelli et al. (1997) ont constaté que la redondance caractérisait 25% des références quand le référent – il s'agissait de champs d'information à remplir – avait un nom très court (une seule lettre). Par contre, elle ne caractérisait que 10% environ des références quand le référent – toujours un champ – comportait un nom plus long (par ex., « Informations générales »). Les auteurs suggèrent l'idée que la redondance serait favorisée quand le coût de la désignation verbale est faible.

Si l'effort est une variable à prendre en compte pour comprendre la redondance multimodale, elle ne semble toutefois pas être la seule. La fiabilité semble aussi jouer un rôle dans la décision d'appliquer deux modalités redondantes. C'est ce qui se dégage de l'analyse des corpus de Catinis (1998) concernant une application de dessin multimodal. Ainsi, l'auteur constate que parmi les actions « dessiner », « déplacer » et « effacer », la dernière est celle qui suscite le plus de redondance. Or, cette action est celle qui comporte le plus de risque (aucune demande de confirmation ne suivait l'action d'effacement). La redondance pourrait donc être exploitée aussi pour fiabiliser une commande.

Usage concurrent de modalités

L'application de modalités concurrentes par l'utilisateur a rarement été observée. Dans l'étude de Mignot et Carbonell (1996), un cas est relaté. L'utilisateur demandait au système de changer l'emplacement du canapé et de l'armoire. Tout en formulant cette commande, il se rendit compte qu'un piano se trouvait sur la trajectoire des meubles et le désigna alors pour le traîner vers le centre de la pièce. Un traitement correct de cette séquence supposerait d'interpréter chaque acte – verbal et gestuel – de façon autonome.

À cette forme de concurrence, on peut ajouter les erreurs typiques d'orientation (ex., désigner la gauche en disant à droite ou l'inverse) ainsi que les lapsus (ex. désigner un objet tout en en nommant un autre par erreur). Toutefois, ces cas devraient être relativement rares, ce qui ne veut pas dire qu'il faudrait en sous-estimer l'importance en particulier si l'erreur peut avoir des conséquences dommageables.

En utilisation réelle, on peut toutefois anticiper qu'un plus grand nombre de cas de concurrence apparaîtra, liés non pas aux actes de l'utilisateur mais aux erreurs de reconnaissance du système. Ainsi, un utilisateur pourrait demander à un système de navigation : « Je voudrais

aller ici », tout en pointant sur la « Rue Minguy » et le système comprendrait « Je voudrais aller Rue du Séjour » par le canal vocal et « Rue Minguy » par le canal gestuel. La conception des systèmes multimodaux devra permettre à l'utilisateur de détecter ces états d'incohérence et les corriger facilement.

Equivalence et assignation des modalités

L'équivalence est rarement analysée et n'a été observée que dans peu d'études. Catinis (1998) fait toutefois état d'un nombre assez important de commandes de dessin (38%) qui sont exprimées par les sujets indifféremment par la parole, avec la souris, ou par une combinaison des deux. Par contre, l'assignation est un phénomène qui semble plus largement répandu.

Il faut souligner que l'assignation peut concerner soit une modalité dans son ensemble – dans ce cas, l'utilisateur a tendance à utiliser le système toujours avec la même modalité quel que soit le contexte de dialogue – soit une modalité (ou combinaison de modalités) pour une commande donnée (Calvet et al., 2001, Catinis, 1988).

L'assignation apparaît moins comme un phénomène s'imposant dès le départ que comme un phénomène émergent d'une succession d'interactions. Certains auteurs ont d'ailleurs noté que cette stabilisation des usages apparaissait relativement vite, pour certains utilisateurs juste après un ou deux essais (Carbonell et al., 1996).

De manière générale, l'assignation semble résulter d'une recherche d'efficacité de l'interaction mais, parfois, elle ne conduit pas forcément l'utilisateur à sélectionner les modes d'interaction les plus efficaces (Mignot et Carbonell, 1996). La question reste toutefois posée quant à savoir si le choix de modalités moins efficaces se fait parce que l'utilisateur y trouve tout de même un intérêt ou s'il se fait par manque de connaissance (voire par oubli) de toutes les possibilités offertes par le système.

Différences inter-individuelles

Plusieurs études font le même constat : la phase d'appropriation d'un système multimodal ne converge pas vers l'adoption des mêmes modalités quel que soit l'utilisateur (Mignot et Carbonell, 1996, Siroux et al., 1997, Calvet et al., 2001 entre autres). Carbonell et al. (1996) parlent de profils ou styles multimodaux individualisés, sans expliquer ce qui conduit les uns ou les autres à préférer telle(s) modalité(s). Calvet & al. (2001) introduisent l'idée de *préférences individuelles* qui inclineraient chaque utilisateur à adopter plutôt telle(s) modalité(s) ou telle(s) autre(s). Ils précisent que ces préférences semblent être liées à de nombreux facteurs, par exemple un goût pour la nouveauté, une aversion pour le vocal (on sait que 15 à 20% des utilisateurs raccrochent immédiatement quand ils s'aperçoivent qu'ils doivent parler à un serveur vocal), ou encore une tendance prononcée pour l'exploration (ou l'inverse). Ces analyses ne conduisent toutefois pas

à penser que l'utilisateur abandonnerait totalement le recours à une modalité non préférée : dans certaines circonstances, on constate en effet une utilisation ponctuelle d'une modalité généralement délaissée.

Changement de modalité face à des erreurs du système

Une situation particulière peut justement provoquer un changement de modalité : il s'agit de l'impossibilité de résoudre rapidement une erreur en gardant la même modalité. Par exemple, Catinis (1998) a fait les constats suivants :

- après une première erreur, l'utilisateur a réitéré la commande avec le même mode dans 78% des cas dans une tâche de dessin et dans 65% des cas dans une tâche de traitement de texte ;
- après le deuxième ou troisième échec, l'utilisateur est passé/e à un autre mode dans 82% des cas pour la tâche de dessin (le calcul n'est pas disponible pour la tâche de traitement de texte).

Parallèlement, plusieurs études ont montré que persévérer en ré-appliquant la modalité vocale après une erreur du système conduisait généralement à dégrader les taux de reconnaissance (ex., Karsenty & Botharel, 2005) et qu'un changement de canal de communication ou de modalité représentationnelle était en général plus efficace pour rétablir la compréhension (Karsenty & Botharel, 2005, Suhm & al., 2001).

La question est de savoir si l'utilisateur novice aurait tendance à changer par lui-même de modalité lorsqu'il rencontre une erreur répétitive ou s'il a besoin d'incitation pour cela. Certaines études tendraient à montrer que, dans ces conditions, le changement de modalité s'effectue « naturellement » (par ex., Oviatt & VanGent, 1996), au moins chez une partie des utilisateurs. Toutefois, d'autres études non publiées dont nous avons eu connaissance ont révélé que les utilisateurs pouvaient persévérer un grand nombre de fois en répétant les mêmes commandes avant de changer de modalité, voire sans jamais le faire. Il est possible que ces observations contradictoires soient liées à des variations dans les conditions d'expérimentation.

DISCUSSION

Les observations passées en revue dans cet article permettent d'identifier 4 grandes classes de déterminants du choix d'une modalité par l'utilisateur :

- la tâche ;
- une fonction d'évaluation subjective de la performance mise en œuvre par l'utilisateur ;
- ses préférences initiales qui le prédisposeraient à opter au départ pour certaines modalités, ou les délaissier, indépendamment de toute autre considération ;
- l'expérience, qui semble pouvoir modifier en partie les préférences initiales de l'utilisateur et concourir à développer, d'une part, l'usage combiné de plusieurs modalités et, d'autre part, l'assignation de

certaines modalités ou combinaisons de modalités à certaines commandes.

Dans cette discussion, nous proposons une analyse de l'effet des deux premiers déterminants afin de mieux en comprendre la nature.

Impact de la nature de la tâche

Plusieurs études invitent à penser que l'usage d'une interface multimodale dépend de la nature de la tâche. Mais comment l'expliquer ? En s'appuyant sur des travaux de Psychologie Cognitive, on peut avancer que la dépendance tâche-modalité repose en premier lieu sur la *nature de la représentation du but* que construit chaque utilisateur. Certaines tâches engagent plus facilement à construire des représentations imagées et spatialisées du but, comme par exemple le déplacement ou la manipulation d'un objet ou l'identification d'un lieu, d'une zone ou d'une distance. D'autres induisent plutôt la formation de représentations dites propositionnelles, comme lors d'une recherche d'un numéro de téléphone ou une requête d'action sur un objet donné (ex., supprimer un fichier). Or, il semble que chacun de ces types de représentation ait un pouvoir de pré-activation de certaines modalités d'action (cf. Krauss et al., 2000) : les représentations imagées pré-activeraient le plus souvent la modalité gestuelle tandis que les représentations propositionnelles pré-activeraient les modalités de production – orale ou écrite - d'un discours. Par ailleurs, il faut noter qu'un geste ne peut représenter que certaines catégories d'information : une forme ou une distance, une localisation dans l'espace, une trajectoire ou une direction. Tous les concepts ne peuvent donc pas être traduits sous forme de geste. A l'inverse, le discours peut théoriquement traduire tout concept mais cette modalité est simplement parfois beaucoup plus coûteuse - on pourrait aussi dire moins « naturelle » - que d'autres pour transmettre certains d'entre eux.

Une autre raison expliquant le lien tâche-modalité est liée aux *exigences à satisfaire dans l'atteinte d'un but*. Certaines tâches imposent par exemple de réaliser le but très rapidement, ou alors avec aucune erreur, ou encore avec un effort minimum. D'autres n'imposent pas ces exigences. Or, chaque modalité ne satisfait pas de la même façon chacune d'entre elles.

Evaluation subjective de la performance

En fait, on peut supposer que l'utilisateur *évalue sa performance avec une modalité donnée* – soit avant de l'utiliser, soit en l'utilisant - pour vérifier qu'elle lui permettra de satisfaire ses exigences de tâche. On peut ainsi rendre compte que l'usage de l'écriture manuscrite sur PDA, qui cause encore un taux d'erreur non négligeable, soit jugée comme acceptable dès lors que l'utilisateur a pour principale contrainte de prendre des notes rapidement. Il est intéressant de noter que les trois exigences que nous avons mentionnées, vitesse d'exécution, effort de mise en œuvre et fiabilité, sont

interdépendantes si bien que lorsque la priorité est mise sur l'une, le poids des autres, ou au moins de l'une des autres, est généralement affaibli. Autrement dit, mettre la priorité sur l'une d'elles doit conduire l'utilisateur à être plus tolérant sur la satisfaction des autres et devrait donc rendre plus acceptable des modalités qui ne les satisferaient pas totalement.

L'évaluation de la performance repose en partie sur une évaluation de l'effort de mise en œuvre avec chaque modalité. A ce titre, on peut comprendre que certaines caractéristiques de l'interface aient un impact sur le choix d'une modalité : c'est le cas par exemple des dispositifs physiques d'utilisation d'une modalité (ex., le Push-To-Talk pour activer la reconnaissance de parole versus l'utilisation libre de la reconnaissance de parole au téléphone) ou du degré de précision exigée dans la réalisation d'un geste. Autre facteur affectant l'évaluation de performance : les caractéristiques de l'objet de l'action. On a notamment vu que la longueur de son nom ou son niveau d'accessibilité à l'écran affectait directement le choix d'une modalité.

Un autre facteur devrait affecter cette évaluation : il s'agit de la modalité sélectionnée pour l'action juste précédente. Catinis (1998) rapporte des travaux ayant montré qu'un changement de modalité entraîne toujours un coût non négligeable. Si l'utilisateur évalue son effort avant d'adopter une modalité donnée dans un contexte donné, ce paramètre doit entrer en jeu et pourrait le conduire à préférer maintenir la modalité active même si elle peut apparaître moins performante lorsqu'elle est considérée isolément. On pourrait expliquer ainsi que certains utilisateurs persévèrent avant de changer de modalité lorsqu'ils rencontrent des erreurs répétées. Cette hypothèse pourrait aussi rendre compte d'une observation inattendue extraite de l'évaluation d'une IHM multimodale sur smartphone (décrite dans Karsenty et al., 2005) : à de nombreuses reprises, les utilisateurs ont confirmé vocalement une requête réalisée en vocal en disant « OK », après avoir activé la reconnaissance de parole en appuyant sur un bouton Push-To-Talk présenté à l'écran, alors que ce même écran présentait un bouton de confirmation « OK », parfaitement visible, sur lequel un simple clic aurait suffi. Dans la très grande majorité des cas, cette confirmation vocale suivait une autre commande vocale.

CONCLUSION

On peut tirer une série d'enseignements de la synthèse des travaux présentée ici et des éléments d'analyse qui ont suivi :

- Pour décider de la pertinence d'une (nouvelle) modalité dans un dispositif d'interaction, les concepteurs devraient prendre en compte en premier lieu la nature même de la tâche et, en particulier, la nature des représentations de but qu'elles induisent chez les utilisateurs ainsi que les propriétés des objets sur lesquels ceux-ci souhaitent agir (ex. degré d'accessibilité, longueur de leur nom, etc.).
- Une fois cette analyse menée, on doit considérer qu'il n'y a toutefois pas une modalité qui est préférable aux autres pour tous les utilisateurs, même pour une tâche donnée. Des différences inter-individuelles fortes existent, pouvant remettre en cause le lien qui pourrait sembler naturel entre une tâche et une modalité. Idéalement, toute tâche devrait donc être réalisable par la modalité supposée la plus efficace pour une tâche donnée et au moins une modalité supplémentaire, dont le choix doit dépendre des acquis antérieurs de la population ciblée.
- Au-delà de ces facteurs, le choix d'une modalité résulterait d'une évaluation subjective de la performance faite par l'utilisateur. Nous avons avancé que la performance avec une modalité donnée était évaluée par rapport à 3 critères principaux inter-dépendants, l'effort de mise en œuvre, la vitesse d'exécution et la fiabilité, tout en reconnaissant que l'utilisateur recherchait probablement le meilleur compromis au vu de ses exigences liées à la tâche. Rien n'empêche donc en théorie qu'il choisisse une modalité moins fiable mais plus rapide que les autres si sa principale contrainte est le temps disponible. Par ailleurs, il apparaît que cette évaluation de performance est fortement contextualisée, dépendant tout à la fois de l'état interne de l'utilisateur (ex., phénomène de pré-activation intra-modale liée à la sortie du système ; modalité utilisée précédemment par l'utilisateur), de l'état du système, des propriétés de l'environnement ou encore des autres modalités disponibles – et donc concurrentes – pour la même action. Une conséquence, parmi d'autres, de cette caractéristique est que la conception et l'évaluation des interfaces multimodales accompagnant le lancement de nouveaux produits devraient s'extraire des conditions de laboratoire – prégnantes dans les 15 dernières années – pour s'ancrer maintenant au plus près des contextes d'usage réels amenant leur lot de complexité... mais aussi de représentativité des conditions d'acceptabilité futures.
- Enfin, on pourra retenir que la mise en œuvre parallèle de deux modalités – thème au cœur de nombreux travaux de recherche sur la multimodalité – est peu probable par des utilisateurs novices (au moins, tant que cette forme d'interaction n'est pas devenue très répandue). L'usage monomodal ou, éventuellement, multimodal sous forme d'une séquence de commandes semble bien plus probable. Par contre, la multimodalité synergique peut apparaître avantageuse avec l'expérience pour certaines tâches et satisfaire les attentes des utilisateurs fréquents. Dans ce cas, concernant la coordination geste/parole, les concepteurs doivent rechercher à créer des systèmes flexibles, acceptant différents modes de coordination entre le geste et la

parole avec certaines limites temporelles entre les deux pour pouvoir distinguer une commande multimodale complémentaire d'une séquence de 2 commandes indépendantes réalisées chacune avec une modalité différente.

Au-delà de ces enseignements, les travaux rapportés ici soulèvent quelques questions qu'il reste à traiter. L'une d'elles concerne l'effort d'apprentissage face à une interface multimodale. Peu d'études abordent cette question pourtant centrale dans la perspective de déployer commercialement des produits ou services multimodaux. Or, à l'image des interfaces vocales dont on pensait à tort qu'elles pouvaient être naturelles, c-à-d., utilisable sans apprentissage (Karsenty, 2002), il apparaît relativement évident que les interfaces multimodales nécessiteront aussi un certain apprentissage. Quelle forme devra-t-il prendre ? Quelles propriétés des interfaces multimodales devront être apprises ? Dans quelle mesure des commandes multimodales pourraient être découvertes dans la continuité des (séquences de) commandes monomodales ? Quelles sont les limites d'acceptabilité des utilisateurs face à cette exigence d'apprentissage ? Voici quelques questions parmi d'autres que la recherche sur la multimodalité devra aborder dans un avenir proche.

REFERENCES

- Bernsen N.O. (2002). Multimodality in language and speech systems. In: B. Granström, D. House and I. Karlsson (eds.) *Multimodality in language and speech systems*. Kluwer.
- Butterworth & Hadar, 1989. Gesture, speech, and computational stages: a reply to McNeill. *Psychological Review*, 96, 168-174
- Calvet G. et al. (2001) Etude empirique de l'usage de la multimodalité avec un ordinateur de poche. *Actes d'IHM-HCI 2001* (papier court), Lille-France.
- Catinis L. et Caelen J. (1995) Analyse du comportement multimodal de l'utilisateur humain dans une tâche de dessin. *Actes d'IHM'95*, pp.123-129.
- Catinis L. (1998). *Etude de l'usage de la parole dans les interfaces multimodales*. Thèse de l'Institut National Polytechnique de Grenoble.
- Cohen, P. R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., & Clow, J. (1997). Quickset: Multimodal interaction for distributed applications. *Proceedings of the 5th International Multimedia Conference*, pp. 31-40. ACM Press.
- Cohen, P. R., Johnston, M., McGee, D., Oviatt, S. L., Clow, J., & Smith, I. (1998). The efficiency of multimodal interaction: A case study. *Proceedings of ICSLP'98*, vol.2, pp. 249-252.
- Coutaz J. & Nigay L. (1994) Les propriétés CARE dans les interfaces multimodales. *Actes d'IHM'94*, pp. 7-14, Lille
- Gaiffe B. & Romary L. (1997). Constraints on the Use of Language, Gesture and Speech for Multimodal Dialogues. *Proceedings of the ACL-EACL Workshop on Referring Phenomena in a Multimedia Context and Their Computational Treatment*, Madrid.
- Hauptmann A.G. & McAvinney P. (1993). Gestures with speech for graphic manipulation. *International Journal of Man-Machine Studies*, 38, 231-249.
- Huang, X. et al. (2001). MIPAD: A multimodal interaction prototype. *Proceedings of ICASSP'2001*, Salt Lake City.
- Huls C. and Bos E. (1998). Studies in Full Integration of Language and Action. In H. Bunt, R. Beuna, and T. Borghuis (Eds.) *Multimodal Human-Computer Communication, Systems, Techniques, and Experiments*, Springer.
- Kamp J.F. (1998). Interaction personnes-systèmes embarqués. Etude des modalités et des dispositifs d'interaction. Thèse d'Informatique de l'ENST.
- Karsenty L. (2002) Shifting the design philosophy of spoken natural language dialogue: From invisible to transparent systems. *International Journal of Speech Technology*, 5(2), 147-158.
- Karsenty L. & Botherel V. (2005) Transparency strategies to help users handle system errors. *Speech Communication*, 45, 305-324.
- Karsenty L., Sire S., Causse M., Deherly D. (2005) Quel impact de l'entrée vocale sur la conception graphique d'un service mobile ? *Actes de la Conférence IHM'2005*, ACM Press.
- Krauss R.M., Chen Y. & Gottesman R.F. (2000). Lexical gestures and lexical access: a process model. In: McNeill D. (ed.) *Language and Gesture* (pp. 261-283). Cambridge University Press.
- Lamel L., Bennacef S., Gauvain J.L., Dartigues H. & Temem J.N. (2002) User Evaluation of the MASK Kiosk. *Speech Communication*, 38(1-2), 131-139.
- Martin J.C., Julia L. & Cheyer A. (1998). A theoretical framework for multimodal user studies. *Proc. of the 2d Int. Conf. on Cooperative Multimodal Communication* (pp. 104-110), Tilburg.
- Mignot C. & Carbonell N. (1996). Commande orale et gestuelle : étude empirique. *Technique et Science Informatiques*, 15(10), 1399-1428.
- Nigay L. & Coutaz J. (1996). Espaces de conception des interfaces multimédia et multimodales. *Techniques et Sciences de l'Informatique*, 15(9), 1195-1225.
- Oviatt S. L. & vanGent R. (1996) Error resolution during multimodal human-computer interaction. *Proc. of the Int. Conf. on Spoken Language Processing*, Philadelphia.
- Oviatt S. L., DeAngeli A. & Kuhn K. (1997). Integration and synchronization of input modes during multimodal human-computer interaction. *Proceedings of CHI '97*, ACM Press, pp. 415-422.

- Petrelli D., De Angeli A., Gerbino W., Cassano G. (1997). Referring in Multimodal Systems: The Importance of User Expertise and System Features. *Proceedings of the Workshop on Referring Phenomena in a Multimedia Context and Their Computational Treatment*, Madrid.
- Siroux, J., Guyomard, M., Multon, F. and Remondeau, C. (1995) Oral and gestural activities of the users in the GEORAL system. *Proceedings of the First International Workshop IMMI*, Edinburgh.
- Suhm B., Myers B. & Waibel A. (2001) Multimodal error correction for speech user interfaces. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 8(1), March 2001, 60-98.
- Thalmann D. (2002) Using Virtual Humans for Multimodal Communication in Virtual Reality and Augmented Reality. In: P.C. Yuen, Y.Y. Tang & P. Wang (eds.) *Multimodal interface for human-machine communication*. World Scientific.