INTERACTION HOMME-MACHINE PAR LA PAROLE ET LE GESTE POUR L'ÉDITION DE DOCUMENTS : TAPAGE

Claudie Faure et Luc Julia
URA CNRS 820
Télécom Paris - Dép. SIG
46 rue Barrault
75634 Paris cedex 13
T : (1) 45 81 71 45
e_mail: cfaure@sig.enst.fr

Résumé : Le projet TAPAGE s'inscrit dans le cadre des d'interfaces multimodales pour l'aide à la conception, l'édition et la manipulation de documents textuels et graphiques. Les modes d'interaction retenus sont la parole et les gestes, ce qui implique des modèles pour l'interprétation des signaux qui supportent la communication humaine. L'interface est construite à partir d'un ordinateur à stylo et d'une carte de reconnaissance vocale. Nous présentons les modes gestuels et parlé en nous limitant à ce qui concerne ce type de matériel. TAPAGE est décrit dans sa version actuellement opératoire, illustrée de résultats. La multimodalité apparaît au niveau des commandes et des données. Le formalisme des événements qui définissent le protocole d'interaction est décrit.

Mots clés : interface multimodale, ordinateur à stylo, dessin

Abstract: TAPAGE is a multimodal interface concerned with designing, editing and manipulating text and graphic documents. The selected modes of interaction are speech and gestures. They imply interpretative models for the signals of human communication. The interface is built using a notepad and a speech recognition system. Gestual and speech modes related to pen and voice are presented. The current working version of TAPAGE is described along with results. Data and commands are both generated in a multimodal way in our application. A formal description of the events defining the interaction protocol is given.

Key Words: Multimodal Interface, Notepad, Line Drawing.

1. INTRODUCTION

Le projet TAPAGE a pour but de réaliser un outil d'aide à la conception, l'édition et la manipulation de données textuelles et graphiques. L'interaction entre l'utilisateur et la machine est supportée par une interface multimodale. Les modes d'interaction retenus sont la parole et les gestes, c'est-à-dire des modalités "naturelles" pour la communication humaine, ce qui conduit à parler d'interface anthropomorphique. L'évolution que connait l'informatique accentue l'intérêt porté aux qualités ergonomiques des interfaces dont l'amélioration doit permettre une meilleure pénétration des outils informatiques dans un public toujours plus vaste. Les interfaces anthropomorphiques visent un gain en temps et en mémoire pour l'utilisateur : une réduction du temps nécessaire pour apprendre à se servir d'une machine, peu de commandes à mémoriser. L'introduction de souris et de menus est une étape importante dans cette évolution qui sera poursuivie par les interfaces acceptant les modes naturels de la communication humaine. Mais cette amélioration implique que la machine puisse comprendre les signaux de la communication humaine. Il sera donc d'abord nécessaire d'en construire des modèles d'interprétation. A cela s'ajoute la nécessité de définir des protocoles multimodaux acceptables par l'utilisateur, qui ne réintroduisent pas des contraintes (sous forme de règles de communication difficiles à mémoriser où à mettre en action) que ce type d'interface était sensé supprimer.

L'interface est développée à partir d'un ordinateur à stylo complété d'une carte de reconnaissance de parole. Nous présentons d'abord les modes d'expression les plus pertinents pour ce type de matériel, c'est-à-dire ceux associés au contact du stylo et du papier. La complémentarité des gestes et de la parole sera abordée pour montrer notamment le caractère contextuel du choix des modes d'expression. Les représentations formelles qui sous-tendent les protocoles d'interaction sont introduites avant de décrire la réalisation d'une maquette de démonstration. Cette maquette a pour but d'aider l'édition de TAbleaux par la PArole et le GEste. La multimodalité est présente au niveau des données et au niveau des commandes, les deux pouvant être produites au stylo, à la voix ou au clavier ou encore en mode mixte (voix et geste).

2. LES MODES GESTUELS

La définition d'une typologie des gestes a fait l'objet de nombreuses études dont : [Greimas, 1970 ; Ekman et Friesen, 1972 ; Argentin 1989]. Nous nous intéressons ici à la communication gestuelle indirecte où le message est véhiculé par l'inscription visible que laisse la main munie d'un instrument de traçage sur la surface du papier réactif et à la désignation directe. Nous parlons de différents modes gestuels pour insister sur la possibilité que donne le geste d'utiliser plusieurs formes d'expression :

- L'écriture ou le mode linguistique. Tout en étant un mode de communication gestuel très naturel, son usage dans les interfaces se voit pénalisé par la difficulté à traiter et interpréter l'écriture naturelle, c'est-à-dire le cursif lié. Ce goulet d'étranglement des interfaces gestuelles constitue un pôle d'étude dans le projet TAPAGE qui ne sera pas développé ici [Bennacer et al., Bercu et al., Ménier et Lorette, 1992]. Les produits actuellement commercialisés reconnaissent des suites de caractères isolés, majuscules ou minuscules. La possibilité d'adapter par entraînement le système de reconnaissance à un scripteur permet d'obtenir des résultats d'assez bonne qualité. Mais la nécessité d'écrire en séparant les lettres impose une écriture contrainte qui gène l'utilisateur.
- Le symbolique, on regroupe ici des vocabulaires de signes définis conventionnellement. La différence avec les caractères alphabétiques réside essentiellement dans une signification attachée aux signes, ce qui bien évidemment ne dispense pas d'une analyse contextuelle. On mettra dans cette classe les chiffres, les symboles mathématiques, chimiques, électriques, typographiques. Les signes typographiques de correction recoivent une attention particulière pour ce genre d'interface. Les commandes gestuelles associées aux corrections signalent par leurs

traces graphiques la nature de l'action à effectuer (effacer, inverser ...) et simultanément les objets sur lesquels portent cette action [Morrel-Samuels, 1990].

- Le graphique, la structure spatiale va jouer un rôle prépondérant dans l'interprétation de ces signaux. Les diagrammes en réseaux, les tableaux, les schématisations d'arrangements spatiaux (plans d'appartements, de jardins, maquettes premier jet de pages de journaux ...) et aussi certains langages visuels de programmation sont des exemples de données qui se rapportent au graphique. Des modèles d'interprétation permettant de passer d'un tracé fait à main levée à une version idéale existent pour des domaines très spécifiques comme les schémas électriques ou les organigrammes [Murase et Wakahara, 1986 ; Okazaki et al., 1988] qui correspondent à des données fortement contraintes.

- La désignation, elle repose sur un mimétisme fonctionnel du stylo et de la souris. Dans le matériel utilisé pour TAPAGE, le stylo est géré informatiquement comme une souris. Le clic de localisation est rendu par un posé de stylo, on peut étendre des fonctions comme "attrape", "étire", "déplace" au stylo. La désignation est très importante pour les interfaces multimodales, on développera par la suite la forte complémentarité de la parole et du geste de pointage. La désignation peut se faire directement sur les données (localiser un mot, une figure) ou sur un menu.

La trace visible du geste n'est pas toujours catégorisable hors contexte (un segment de droite dans un dessin et une barre de fraction ont même apparence). On reviendra sur ce point à propos du protocole d'interaction.

La catégorisation ci-dessus reflète les différents "modules" d'analyse automatique des signaux gestuels enregistrés. Ces modules se différencient surtout par la nature des connaissances mises en jeu pour reconnaître ou interpréter ces traces graphiques et par la nature des informations à extraire des signaux. On note que les modes gestuels peuvent être catégorisés sur d'autres critères : suivant que le geste produise ou non une trace (ce qui sépare la désignation des autres classes citées), ou suivant la fonction du geste (production d'une commande ou d'un élément visible du document élaboré).

3. GESTES ET PAROLE

Le geste et la parole appartiennent à un même programme d'émission d'un message. A sa réception, la compréhension s'appuiera sur les caractéristiques physiques et le contenu linguistique du signal de parole ainsi que sur les signaux gestuels (gestes des mains, mouvements des lèvres, attitudes du corps ...). Nous nous intéressons ici, dans le cadre des interfaces homme-machine multimodales, à la complémentarité du geste et de la parole en ne retenant que les gestes pratiques, intentionnellement produits dans le but de compléter ou de préciser un message parlé. L'usage du mode parlé ou gestuel est déterminé par une recherche d'économie qui porte sur plusieurs facteurs dont nous retenons : la rapidité, la précision, la simplicité. La rapidité concerne le temps d'émission (de réception) du message, la précision vise à produire des messages non-ambigus et la simplicité d'un message relève de la complexité des mécanismes nécessaires à la construction (à l'analyse) de son expression.

L'exemple typique que l'on donne souvent de la complémentarité économique du geste et de la parole est :

(C1) /met ça là/ accompagné d'une désignation spatiale du /çà/ et du /là/.

Cet exemple est aussi typique des problèmes de gestion d'événements dans une interface multimodale. Dans le cas (C1), l'intention de déplacement est rendu par un verbe d'action (/met/) qui implique un objet. Il est présent dans la séquence parlé (/ça/) mais non spécifié. Le démonstratif économise une dénomination précise, voire des constructions linguistiques qui permettraient de l'identifier, alors qu'un seul geste renvoit à l'objet physique de référence. De même la position (/là/) évite de décrire verbalement une position spatiale désignée par un geste.

Plus généralement, l'information spatiale est difficile à traduire avec précision dans un mode d'expression linguistique. /A est sur B/ signifie que A est posé sur B, mais /A est au-dessus de B/ ne donne qu'une plage de positions possibles pour A associées à des valeurs de choix variables en fonction de la distance de A à B, des tailles relatives de A et B, de la taille de l'espace disponible, etc. Pour des informations spatiales portant sur des grandeurs, le discours spontané est souvent accompagné de gestes qui miment l'action (/agrandir/, /resserrer/) ou spécifient les tailles (/grand comme ça/).

Le geste sera donc plus efficace pour exprimer des caractéristiques spatiales quantitatives comme la position d'un point dans un espace ou la taille d'un objet (dans une commande comme /agrandir/).

Même après catégorisation des types d'information, on ne peut pas conclure à la supériorité d'un mode sur un autre. On trouvera dans [Kurtenbach et Hulteen, Mountford et Gaver, 1990] une discussion sur ces modes d'interaction. Le meilleur choix, au sens d'une économie, est toujours contextuel. On donne quelques exemples qui illustrent la relation entre le choix des commandes et le contexte visuel auquel elles se rapportent.

Dans une image graphique composée d'un ensemble de figures géométriques on peut vouloir effacer une ou plusieurs de ces figures. On considère une suite de commandes :

(C2) /efface çà/ et geste
(C3) /efface les carrés/
(C4) /efface les figures de gauche/
(C5) /efface les figures/ et gestes

La commande (C2), conformément à (C1), permet d'effacer un objet désigné par un geste. L'économie est du type de celle décrite pour (C1). La commande (C3) permet d'effacer les carrés présents sur l'image, le pluriel et l'attribut "carré" spécifique des figures à éliminer évite les gestes répétitifs de désignation qui seraient nécessaires. La commande (C4) permet de désigner un groupement perceptif et de le traiter comme un seul objet. Pour (C5) on peut considérer que les figures à effacer ne sont pas identifiables par un simple attribut ou une combinaison de quelques attributs (de forme, de couleur, de taille), qu'elles sont réparties dans l'espace de telle manière qu'aucun groupement perceptif n'apparaît ou que, s'il apparaît, la verbalisation de sa position pose problème (ni le plus à gauche, ni le plus haut ...). La nature des gestes sera donc fonction de la répartition des figures dans l'espace. Un groupement perceptif fait apparaître un objet virtuel localisé qu'un seul geste peut désigner par encerclement de ses composantes. Dans ce cas, (C2) peut être utilisée pour effacer le groupe d'objets. En l'absence de cette structuration spatiale, les gestes pointeront sur chaque objet à effacer.

La multimodalité permet d'introduire des menus parlés spécifiant le type d'action (/lier/, /effacer/ ...), le geste désignant les objets et les positions sur l'espace de travail. L'utilisation du seul mode gestuel pour intervenir à la fois sur les menus et l'espace de travail de l'écran oblige à déplacer l'attention gestuelle et visuelle de l'espace des données (où le geste spécifie de lieu de l'action) au menu (où le geste déclenche une action).

Les actions verbalisées peuvent être associées à des objets graphiques génériques. Nommer l'action permet à la fois de lancer son exécution et d'appeler les objets qui permettent de la réaliser. Par exemple /lier/ peut se ramener à appeler une procédure de traçage de segments (les objets graphiques attachés au verbe d'action) entre des figures désignées. Cette situation est identique à la numérotation des pages : par la commande /numéroter les pages/, les objets numéros ne sont pas manipulés directement mais par l'intermédiaire de la commande.

4. PROTOCOLE D' INTERACTION

Les modèles d'architectures logicielles d'interfaces homme-machine ne permettent pas de définir de manière opératoire des interfaces multimodales. La spécificité du multimodal soulignée à propos des "langages" de commande [] et qui pose

essentiellement le problème de la co-référence [] et des modes d'intégration des différents signaux du dialogue, doit trouver sa manifestation au niveau de l'architecture comme le souligne [Vigouroux et al., 1989]. Nous retenons la notion d'agents, sur laquelle s'appuie le modèle PAC [Coutaz, 1990], elle permet de transposer les caractéristiques du dialogue multimodal comme le parallélisme et la fusion en terme d'autonomie et de coopération des agents. L'architecture proposée possède trois composantes principales :
- Des agents de présentation permettent de saisir en parallèle les gestes et les signaux de parole émis par l'utilisateur. Ce parallélisme confère aux commandes une qualité "synergique" au sens défini par [Bellik et Teil, 1992]. Ces agents agissent sur les sorties, de manière autonome, en retournant un signal à l'utilisateur (l'objet pointé change d'aspect, par exemple). Cette caractéristique sera exploitée pour assurer une multimodalité en sortie.
- L'interprétation des informations saisies est guidée par le verbe détecté : les analyses syntaxique et sémantique sont indissociables. La signification du verbe guide la recherche des informations qui peuvent être nécessaires pour compléter la commande. Ces informations sont recherchées dans les mémoires des agents de présentation où elles ont été enregistrées dans un ordre indéfini.
- Un agent de contrôle du dialogue active l'exécution de la commande quand sa forme complète est obtenue.

Les protocoles d'interaction sont définis à partir d'événements qui peuvent être catégorisés comme donnée ou commande. Les commandes se divisent en commandes d'action qui transforment l'apparence des données visualisées, et en commandes d'état qui spécifient des modes ou des types d'actions. La figure 1 donne une description des événements. Les commandes sont indépendantes les unes des autres, la détermination des références n'implique donc pas une analyse de l'historique mais par contre elle implique de résoudre des problèmes de co-références propres au multimodal.

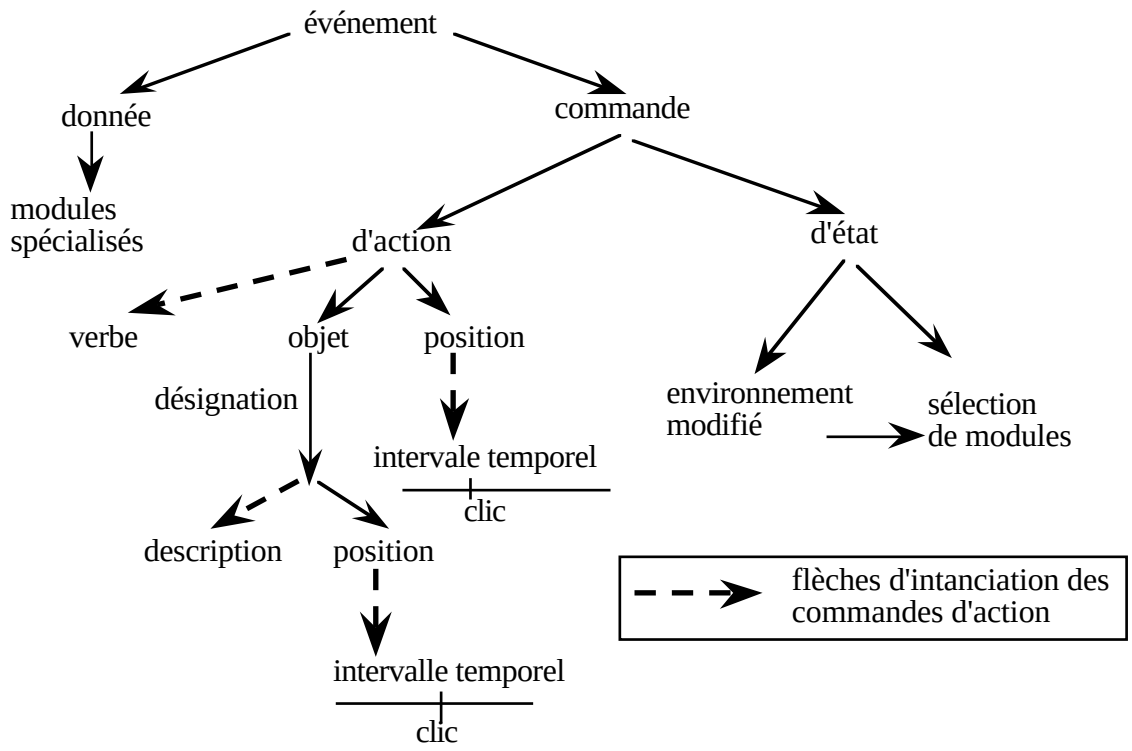

Figure 1
Schéma des relations entre le niveau abstrait d'événement et les entités physiques

On décrit d'abord la forme de représentation d'une commande d'action en commençant par le niveau le plus abstrait. A ce niveau l'action est définie comme un

ensemble d'unités d'information (UI). Ce niveau de description est indépendant des moyens mis en œuvre pour instancier les UI, c'est-à-dire indépendant des modes d'interaction. On retient pour l'instant une forme où l'action se présente comme un triplet d'UI : <verbe objet position >. Cette forme ne peut pas couvrir l'ensemble des commandes que nous envisageons, en particulier celles qui nécessitent de préciser des variables pour spécifier complètement des actions du type /colorier/, /agrandir/. Elle peut néanmoins supporter de nombreuses commandes d'action :

	verbe	objet	position
1.	/efface/	/le/	
2.	/met /	/ça/	/là/
3.	/dessine/	/un carré/	/au milieu/
4.	/aligne /	/les cercles/	/à gauche/

Les niveaux de représentation subordonnés permettent de spécifier complètement l'énoncé de la commande en instanciant les UI. Ils sont définis de manière à pouvoir intégrer plusieurs modes d'interaction de manière transparente pour l'utilisateur. Dans les exemples 1 et 2, la commande verbale n'est pas complète. L'objet de référence du /le/ et du /ça/, de même que la position de référence du /là/ sont donnés par le contexte. Les représentations des objets et des positions doivent permettre de compléter ces formes verbales par des informations tirées du contexte afin d'obtenir l'instanciation des UI correspondantes. Un objet et une position sont attachés à des champs de désignation qui sont remplis dynamiquement par des valeurs spécifiques à un seul objet présent sur l'écran de visualisation ou à un seul lieu. Dans le cas de la position, le champ de désignation est un intervalle de temps pendant lequel se produit une commande gestuelle : le pointage sur l'écran. Dans le cas d'un objet, on dispose de deux champs de désignation possibles. Le premier est une position spécifiée comme ci dessus par une désignation gestuelle dans un intervalle de temps, le deuxième est un attribut descriptif spécifiant la forme (ex : /le carré/), la position spatiale (ex : /la figure du haut/), ou d'autres expressions descriptives non ambiguës dans le contexte de visualisation (ex : /les cercles grisés/). On peut donc spécifier un objet par une description discriminante ou par son emplacement dans l'espace de visualisation.

Un analyseur interprète la commande d'action en associant au verbe les instances d'UI nécessaires pour effectuer l'action. Cette forme de représentation s'adapte parfaitement à plusieurs modes d'interaction sans qu'il soit besoin de contraindre l'utilisateur à un choix a priori. La forme figée de la syntaxe d'action est reportée au niveau le plus abstrait pour permettre une souplesse au niveau de la saisie des informations qui instancient complètement la commande d'action.

Les commandes d'état sont effectuées à partir de menus visualisés. Les états peuvent être choisis à la voix ou par désignation gestuelle. Ces états spécifient des modes d'interaction, comme le dessin ou l'écriture, ou des actions génériques, comme la correction. La spécification des états "dessin" et "écriture" est nécessaire pour lever les ambiguïtés de reconnaissance sur des tracés pouvant être interprétés comme des figures géométriques ou des caractères alphabétiques (un cercle et un O par exemple). Ambiguïté d'autant plus grande que la reconnaissance de l'écriture se fait à mesure que les tracés sont produits et qu'il ne s'agit pas de supprimer cette reconnaissance rapide.

Une commande d'état sélectionne les modules de traitement spécifiques à l'état (ils sont alors activables après réception des données) et désactive d'autres modules (la reconnaissance de caractères est désactivée en mode dessin). Une telle commande peut aussi transformer l'environnement visuel en affichant des menus spécifiques à certains états (comme l'état correction) qui correspondent à différentes étapes dans la conception d'un document. La possibilité de nommer verbalement des états du menu permet à l'utilisateur de ne pas déplacer son attention visuelle des données au menu.

5. LA RÉALISATION

Une maquette de démonstration sur une application adaptée à la technologie et à la vocation des ordinateurs à stylo est réalisée. Cette maquette est évolutive, elle

progresse en fonction des modules de traitement intégrés et des architectures qui définissent les protocoles d'interaction multimodale. On distingue la conception de l'interface de celle du noyau fonctionnel qui contient les modules de traitement des différents signaux d'entrée. Dans une première version, sa fonction consiste à idéaliser automatiquement des dessins manuels de tableaux, à les remplir de chiffres ou de mots, à les corriger. Les modes d'interaction sont la parole, l'écrit (stylo et clavier), le dessin, le pointage et les commandes gestuelles (pointage, symboles).

5.1. L'idéalisation des tableaux

Ce module du noyau fonctionnel comprend plusieurs étapes de traitement que nous présentons brièvement. Les signaux se présentent sous forme de suite de points (figure 2). Dans une première étape l'ensemble des tracés est découpé de manière à trouver un ensemble de verticales (V) et d'horizontales (H) qui l'approxime. Les tracés sont d'abord polygonalisés par une méthode de type *SPLIT and MERGE* [Pavlidis, 1982] qui les découpe en segments V et H compte tenu d'un paramètre de tolérance, puis regroupe les segments de même direction sur un critère de voisinage. Les segments obtenus sont alors redressés suivant les directions H et V. Dans une deuxième étape, les jonctions en L et T sont détectées et reconstruites idéalement (prolongement des segments pour une mise en contact effective au niveau des points de jonction, effacement des segments parasites).

La troisième étape concerne un niveau de traitement qui s'appuie sur les valeurs numériques des données et sur une connaissance du domaine. Les tableaux sont des structures spatiales qui suivent les principes de la communication visuelle dont celui qui veut que les différences physiques portant sur un attribut (taille, forme, couleur …) signalent des différences sur le plan de l'interprétation [Bertin, 1977]. La communication visuelle tend à éviter les différenciations qui introduiraient des informations non souhaitées [Marks et Reiter, 1990]. Des groupes de colonnes, de lignes et de cases ont des tailles identiques dans un tableau. Or le dessin manuscrit ne produit jamais des grandeurs exactement égales. Cette égalisation résultera d'une reconstruction automatique, d'autant plus difficile à faire que les sous-ensembles de colonnes, de lignes et de cases égales ne sont pas connus a priori. Une recherche dynamique de motifs répétitifs est effectuée en parcourant d'abord l'axe V puis l'axe H par un algorithme récursif. Les séparateurs de motifs sont fixés chaque fois qu'une séquence de motifs est détectée. Les zones de recherche des motifs et l'ordre dans lequel elles sont examinées ne sont pas définis a priori : ils dépendent de la position des segments fixes qui résultent des étapes précédentes de l'analyse. Des motifs ayant des tailles égales, compte tenu du paramètre de tolérance, sont reconstruits strictement égaux. La figure 3 illustre le résultat obtenu.

Figure 2
Tableau manuscrit

Figure 3
Après idéalisation et remplissage d'une case

Contrairement à des problèmes de reconnaissance de formes où les signaux traités appartiennent à des classes (de sons, de mots ...) connues a priori et en nombre fini, un tableau n'est pas "reconnu" comme élément d'une classe. La tâche du scripteur consiste à dessiner ses tableaux et non à recopier des tableaux types. On ne peut donc pas évaluer les performances du traitement par un pourcentage de bonnes assignations des tableaux à des classes.

Les différences de style des scripteurs portent sur la variabilité au niveau des éléments constituants du tableau (lignes, angles, fermeture, motifs ...), elles sont prises en compte par des seuils de tolérance dans les différentes opérations de reconstruction. Nous illustrons la robustesse de la méthode (figures 2 et 3) par un exemple de tableau fortement "bruité" (lignes fluctuantes, plusieurs tracés pour un seul segment, jonctions et angles imparfaits ...) et de structure complexe présentant plusieurs motifs (différentes tailles de cellules). Le style toléré ne contraint pas le scripteur à une grande application dans son tracé. On note que la variabilité du style semble directement lié à la vitesse de production.

Il faut aussi considérer que les erreurs sont moins dommageables dans un système où le scripteur agit directement sur les données, notamment en utilisant la correction pour réparer les erreurs du traitement, et ceci par opposition aux systèmes de traitement n'offrant pas cette interaction directe. Dans cette application, nous nous situons dans le cadre d'une tâche de conception de dessins de sorte que la correction n'a pas pour seul but de réparer les erreurs du traitement mais de permettre au scripteur d'élaborer son dessin en le faisant évoluer.

5.2. La multimodalité

TAPAGE est développé à partir d'un ordinateur sans clavier du commerce muni d'un système de reconnaissance de caractères manuscrits. Après apprentissage par un scripteur (plusieurs heures) la reconnaissance donne de très bons résultats pour ce scripteur. Le stylo est sans fil, sa gestion informatique est assimilable à celle d'une souris. Nous avons d'abord adapté ce matériel à notre application pour pouvoir saisir de l'écriture et des dessins. La reconnaissance de la parole est faite par une carte DATAVOX commercialisée par VECSYS. Le vocabulaire est limité pour l'instant à 24 mots de base et 27 synonymes Les mots sont prononcés isolément ou en continu, la reconnaissance est monolocuteur mais accepte des locuteurs voisins.

L'utilisateur dispose d'un écran sur lequel est affiché une colonne de boutons, d'un stylo, d'un micro et éventuellement d'un clavier connecté ; il peut aussi faire apparaître un clavier virtuel ou une fenêtre d'écriture par des commandes gestuelles spécifiques. Il choisi d'afficher les boutons à gauche ou à droite de l'écran, il dispose ainsi d'un menu. Le menu principal correspond à des commandes indiquant un état. Selon le bouton sélectionné, un menu contextuel peut apparaître : après sélection de l’état correction le menu affiché contient des boutons permettant des actions comme le déplacement et l'effacement. Les boutons sont activables par deux modes de communication : la voix (le nom inscrit sur le bouton est prononcé) ou le geste (le stylo pointe sur le bouton). Ces deux modes sont toujours possibles et laissés au choix de l'utilisateur à chaque activation d'un bouton. Une fois sélectionné, le bouton s'enfonce ce qui se traduit visuellement par un changement d'aspect, l'utilisateur peut ainsi vérifier à tout moment dans quel contexte il opère.

La figure 4 illustre les opérations permettant d'obtenir le résultat de la figure 3. Dans l'action qui correspond à la saisie des données graphiques, le verbe est instancié par /dessine/ prononcé ou pointé sur le menu. La trace du stylo spécifie à la fois l'objet (le dessin en cours de réalisation) et la position (la pointe du stylo). L'écriture correspond au verbe /écris/ prononcé ou pointé sur le menu, la position est désignée par pointage, dans cette application il s'agit du pointage d'une case du tableau qui une fois sélectionnée change d'aspect pour conforter l'utilisateur. Le contenu de la case est l'objet de la commande d'écriture. Il peut être rentré au clavier connecté, ou au clavier virtuel ou au stylo. Ces deux dernières possibilités nécessitent de faire apparaître le clavier virtuel ou la fenêtre d'écriture par des commandes gestuelles. L'ordre dans

lequel la case et son contenu sont spécifiés n'est pas fixé. On remarque que des commandes d'état instancient de manière transparente pour l'utilisateur le verbe d'une commande d'action.

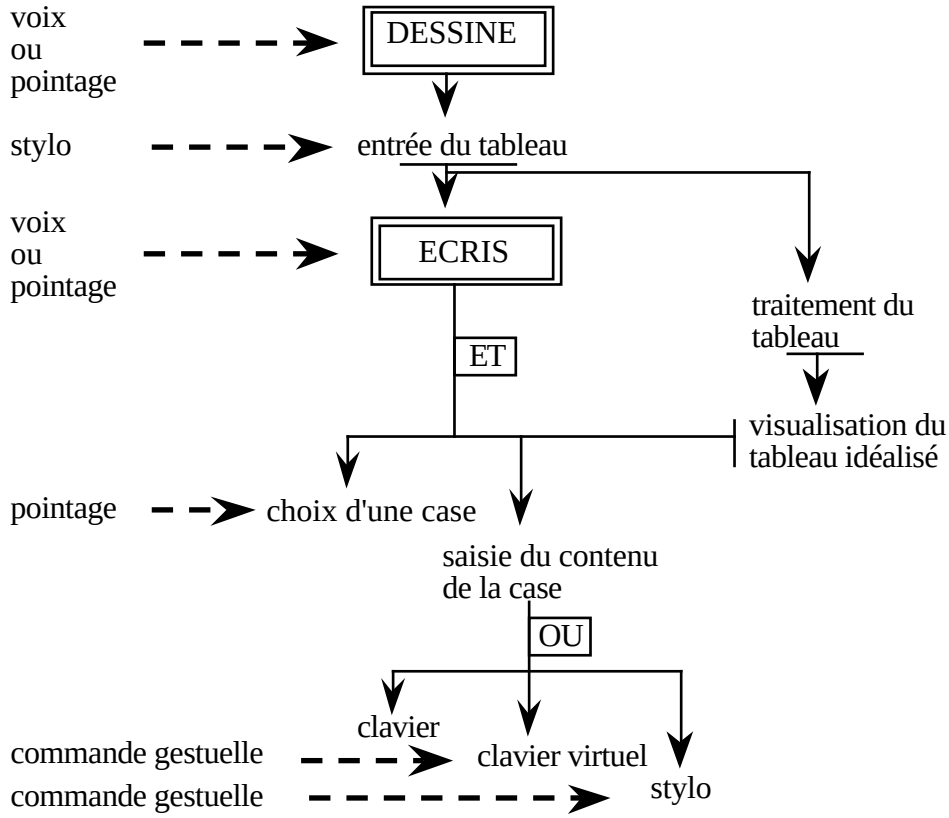

Figure 4
Les opérations conduisant au résultat de la figure 3

L'ensemble de ces commandes intègrent différents modes d'expression qui ne sont pas fixés à priori, réalisant ainsi une véritable synergie des modes. Pour résumer les rôles attribués à la voix et au geste :
- la voix permet de spécifier des états et par suite des verbes d'action, et de désigner des objets par des descriptions,
- le stylo permet de dessiner, d'écrire, de pointer des objets ou des positions ainsi que des boutons du menu affiché, de produire des commandes gestuelles faisant apparaître des outils de saisie.

6. COMMENTAIRES

L'objectif des interfaces anthropomorphiques est de proposer à l'utilisateur un outil informatique qui lui donne une impression de familiarité et de simplicité. Les protocoles d'interaction que nous avons définis et réalisés devront être testés en situation d'utilisation. Cette pratique expérimentale nous semble être le meilleur moyen pour en faire une véritable évaluation du point de vue ergonomique, mais aussi pour atteindre de meilleures définitions de protocoles. Il faut néanmoins garder à l'esprit que le matériel impose ses propres contraintes. Le barrage qui reste le plus délicat à passer est celui de l'interprétation des signaux de la communication humaine. Nous avons déjà signalé le problème que pose la reconnaissance d'une écriture manuscrite naturelle, la reconnaissance de la parole n'est pas suffisamment fiable pour satisfaire un utilisateur quelles que soient les conditions de l'environnement (milieu bruité, micro mal positionné ...). Pour le graphique (tableau, schémas explicatifs ...) le passage des données

manuelles imparfaites à une version idéale nécessite des modèles d'interprétations qui restent à préciser.

Ces remarques montrent la nécessité de travailler à la fois sur le plan informatique et sur le plan humain d'une part pour définir les protocoles et d'autre part pour construire des modèles d'interprétation des signaux de la communication humaine.

Nous nous sommes restreint au triplet <action objet position> pour décrire les protocoles d'interaction. Pour exprimer naturellement des commandes, des formes plus riches peuvent être nécessaires et par suite des syntaxes plus élaborées. La définition de ces formes de commandes doit se faire dans le contexte du multimodal où la possibilité d'utiliser la désignation gestuelle évite de recourir à des énoncés aux formes linguistiques complexes mais implique une forte complémentarité du verbal et du gestuel dans les énoncés, ce qui amène à définir des syntaxes multimodales.

Références

ARGENTIN G. (1989) *Quand faire c'est dire*. Pierre Mardana, éditeur.

BENNACER L., LEMOINE J., PETIT E. (1992) Une méthode en ligne de reconnaissance d'écriture par double balayage. *Actes de CNED'92, BIGRE n° 80.* pp. 333- 338.

BERCU S., DELYON B., LORETTE G. (1992) Segmentation pour une méthode de reconnaissance d'écriture cursive en-ligne. *Actes de CNED'92, BIGRE n° 80.* pp. 144-151.

BERTIN J. (1977) *La graphique*. Flammarion.

EKMAN P., FRIESEN W.V. (1972) Hand movements. *The Journal of Communication.* pp. 353-374.

GREIMAS A.J. (1970) *Du sens*. Editions du Seuil.

KURTENBACH G., HULTEEN E.A. (1990) Gestures in Human-Computer Communication. *The art of human-computer interface design*. Adison Wesley, pp. 309-317.

MARKS J., REITER E. (1990) Avoiding Unwanted Conversational Implicatures in Text and Graphics. *Proc. Eight National Conference on Artificial Intelligence*, The MIT Press, pp. 450-456.

MENIER G., LORETTE G. (1992) Segmentation et reconnaissance en ligne d'écriture cursive à l'aide de plusieurs niveaux d'information contextuelle. *Actes de CNED'92, BIGRE n° 80*, pp. 318-332.

MORREL-SAMUELS P. (1990) Clarifying the distinction between lexical and gestural commands. *International Journal of Man-Machine Studies*, 32, pp. 581-590.

MOUNTFORD S.J., GAVER W.W. (1990) Talking and Listening to Computers. *The art of human-computer interface design*. Adison Wesley, pp. 319-334.

MURASE H., WAKAHARA T. (1986) On-line Hand-Sketched Figure Recognition. *Pattern Recognition* Vol. 19, N°2, pp. 147-160.

OKAZAKI A., KONDO T., MORI K., TSUNEKANA S., KAWAMOTO E. (1988) An Automatic Circuit Diagram Reader with Loop-Structure Based Symbol Recognition. *IEEE PAMI* Vol. 10, N°3, pp. 331-340.

PAVLIDIS T. (1982) *Structural Pattern Recognition*. Springer Verlag.