THESE

présentée par

Annelies BRAFFORT

pour l'obtention du titre de DOCTEUR de l'Université de Paris-XI
Spécialité : INFORMATIQUE

Reconnaissance et compréhension de gestes, application à la langue des signes

Date de soutenance : 28 Juin 1996

Composition du jury :

Président :

M.	**G. Ligozat**	**(U. Paris XI - LIMSI)**

Rapporteurs :

Mme.	**J. Cassell**	**(Medialab MIT, USA)**
M.	**C. Cuxac**	**(U. René Descartes, Paris)**
Mme.	**C. Faure**	**(ENST, Paris)**

Examinateurs :

M.	**J. Mariani**	**(LIMSI)**
M.	**D. Teil**	**(LIMSI)**

Thèse préparée au sein du
Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
LIMSI-CNRS

Remerciements et dédicaces

Je tiens à remercier ici toutes les personnes qui m'ont aidée à relever le gant pour l'avoir bien en main, en particulier :

Daniel Teil, qui d'une main de fer dans un gant de velours m'a laissé les mains libres,

Françoise Forest, qui d'un signe de la main m'a mis sur le che-min,

Sylvie Gibet, qui d'une main experte m'a prêté main forte,

Joseph Mariani, qui a su de main de maître diriger la main de l'"artiste",

Christian Cuxac, pour avoir pu faire main basse sur tout ce qui grâce à lui m'est tombé sous la main,

Justine Cassell, qui a mis la main à la pâte et à travers l'atlantique m'a tendu la main,

Claudie Faure, dont la main sûre a soutenu mon geste,

Gérard Ligozat, qui a su trouver le temps et l'espace à la portée de sa main pour lire mes lignes (de main),

et mes collègues, amis et plus si affinité, dont Angel, Anne, Boris, Christophe, David, Édouard, François, Gérard, Guy, Mike, Patrick, Rachid, Thierry, Xavier, Yacine, et tous les autres, sans oublier tutta la mia famiglia, pour, entre autres :

leur main d'oeuvre, leurs coups de main, leurs mains secourables, leurs mains lestes et heureuses, leurs informations de première main,

qui m'ont permis de me faire la main, d'avoir tout sous la main, de ne pas perdre la main et de ne jamais avoir les mains vides.

Même si nous avons parfois été pris la main dans le sac à lever le coude sans prendre de gant ni y aller de main morte, nous n'en sommes jamais venus aux mains et, la main dans la main, avons fait mentir le vieil adage "jeux de mains, jeux de vilains".

J'ai peut-être eu la main un peu lourde ici, non ?

J'en mettrais ma main à couper

…

Note

Tous les dessins illustrant les signes de la LSF présentés dans cette thèse sont issus des livres de Bill Moody (La langue des signes, tomes 1 et 2, Ellipses).

RESUME

Cette thèse s'inscrit dans le domaine de la communication homme-machine et plus spécifiquement dans celui de la communication gestuelle. L'objectif est la conception de systèmes de reconnaissance et de compréhension adaptés au canal gestuel afin de pouvoir intégrer de nouvelles possibilités d'interactions au sein d'interfaces informatiques. Une attention particulière a été portée au problème de la reconnaissance et de la compréhension de la langue des signes française (LSF).

Une étude détaillée de la structure et du fonctionnement de la LSF a été menée en vue de sa modélisation informatique. Parmi les différentes propriétés dégagées, les deux suivantes sont particulièrement importantes. Tout d'abord, un signe est multidimensionnel car il possède la propriété de transmettre plusieurs types d'informations simultanément, par l'intermédiaire de paramètres qui sont au nombre de quatre pour les gestes de la main (configuration, mouvement, orientation et emplacement). Par ailleurs, deux catégories de signes doivent être distinguées : les gestes pour lesquels les quatre paramètres sont invariables quelque soit le contexte (les signes standard) et ceux pour lesquels au moins un des paramètres est variable en fonction du contexte (les signes variables).

Le système proposé, nommé ARGo, tient compte des résultats de cette étude. Il est composé d'un module de reconnaissance et d'un module de compréhension. Le module de reconnaissance est basé sur une approche markovienne qui a permis l'obtention de résultats très encourageants sur des phrases de gestes enchaînés composées de signes standard et variables (taux de reconnaissance 96%). Le module de compréhension est fondé sur la définition de règles spatio-temporelle et la représentation de la scène de narration du signeur. Un prototype comportant une scène virtuelle permet de visualiser étape par étape le processus d'interprétation. A la fin du processus, une traduction de la phrase en français est produite.

MOTS-CLEFS

Communication Homme-Machine, Interaction Gestuelle, Langue des Signes Française, Reconnaissance de Gestes Enchaînés, Compréhension de Phrases Gestuelles.

ABSTRACT

This thesis lies within the scope of human-machine communication, and more specifically within gestural communication. The aim is to design recognition and understanding systems dedicated to the gestural channel, in order to integrate new kinds of interaction inside computer interfaces. A particular attention has been put on the French Sign Language (LSF) recognition and understanding problem.

A detailed study of the LSF structure and functioning has been carried out in order to build a computing model. Among the different extracted properties, the two following are particularly important. First, a sign is multi-dimensional, because it has the property of conveying several kinds of information simultaneously, by means of four parameters for the hand gesture (configuration, movement, orientation and location). Futhermore, two categories of signs must be distinguished: those for which the four parameters are invariable for any context (the standard signs) and those for which at least one parameter is variable according to the context (the variable signs).

The proposed system, named ARGo, takes into account the results of this study. It is composed of a recognition module and an understanding module. The recognition module is based on a Markov approach, which has provided very encouraging results on linked gesture sentences composed of standard and variable signs (recognition rate 96%). The understanding module is based on the definition of spatio-temporal rules and the representation of the signer narration scene. A prototype including a virtual scene allows us to visualize step by step the interpretation process. At the end of the process, a translation of the sentence in French is produced.

KEY-WORDS

Human-Machine Communication, Gestural Interaction, French Sign Language, Linked Gesture Recognition, Gestural Sentence Understanding.

Sommaire

Introduction **1**

1. La communication gestuelle **5**

1.1. Positionnement du problème *6*

1.2. Le geste de commande *12*

1.3. Le geste co-verbal *26*

1.4. Le geste de la langue des Signes *39*

1.5. Conclusion *48*

2. Étude des paramètres de la LSF **49**

2.1. Introduction *50*

2.2. Définition des paramètres *51*

2.3. Les quatre paramètres *53*

2.4. Autres informations *89*

2.5. La base de données *91*

2.6. Conclusion *94*

3. La reconnaissance de gestes **95**

3.1. État de l'art *96*

3.2. Outils utilisés et développés *126*

3.3. Expérimentations réalisées *136*

3.4. Conclusion *162*

4. Vers un système de compréhension **163**

4.1. Modélisation de la scène de narration *164*

4.2. Architecture du système ARGo *170*

4.3. Fonctionnement : un exemple *178*

4.4. Évaluation *185*

4.5. Applications *186*

4.6. Conclusion *191*

Conclusion et perspectives **193**

Bibliographie **197**

Annexes **211**

Table des matières **225**

INTRODUCTION

"*Pourquoi, sans la connaître [la langue des signes], la rejeter dans un coin, d'une main dédaigneuse, comme un instrument inutile ?*

Pourquoi ne pas l'étudier plutôt ?…"

Berthier, *Observations sur la mimique*, 1853

Cette thèse porte sur l'étude de la modalité gestuelle et son utilisation au sein de la communication homme-machine. Selon l'angle sous lequel on regarde le problème, deux questions peuvent se poser :

- Du point de vue informatique, que peuvent être les apports de la modalité gestuelle à la communication homme-machine ?
- Du point de vue humain, que peut apporter une interface gestuelle aux personnes qui utilisent un ordinateur ?

La première question pose le problème de l'intégration de la modalité gestuelle au sein d'applications informatiques. L'apparition de nouveaux dispositifs de capture du geste (gant numérique, capteur de position, oculomètre…) a favorisé l'émergence de nouveau thèmes de recherche visant à intégrer la modalité gestuelle au sein d'applications informatiques. Ce canal d'information apporte alors de nouvelles possibilités d'interactions. Cependant, il ne suffit pas d'ajouter ces nouveaux systèmes aux dispositifs classiques tels que le clavier, ou la souris. Des problèmes spécifiques se posent, tels que la conception d'algorithmes de reconnaissance adaptés et la mise en oeuvre de systèmes de compréhension des gestes.

La seconde question pose le problème des applications possibles pour les utilisateurs d'ordinateurs en général et pour les personnes pour lesquelles la modalité gestuelle est prédominante en particulier.

- Le fait d'ajouter un canal de communication gestuel à un ordinateur doit permettre d'offrir à l'utilisateur des nouveaux types d'interfaces mieux adaptés pour certaines applications. Par exemple, si l'application consiste à décrire une scène dans l'espace à des fins de modélisation d'objets ou de scènes virtuelles, de description d'itinéraires, ou d'autres types d'applications pour lesquels des informations spatio-

temporelles doivent être exprimées, la modalité gestuelle sera mieux adaptée que le langage naturel. Elle peut être intégrée au sein d'une interface multimodale permettant à l'utilisateur de bénéficier d'interactions plus riches avec l'application.

- Les personnes pour lesquelles la modalité gestuelle est prédominante sont typiquement les personnes sourdes. Un des problèmes cruciaux de la communauté Sourde en France est lié à l'éducation. Du fait que la langue des signes a été interdite en France pendant plus d'un siècle (1880-1991), la langue utilisée pour enseigner les connaissances aux enfants sourds est souvent le français. Il est admis dans beaucoup de pays que cet état de fait est source de nombreux échecs scolaires. Le développement du Minitel et la démocratisation des ordinateurs portables ont permis aux sourds d'accéder à de nouveaux types d'outils leur permettant de communiquer entre eux plus facilement et d'accéder à de nouveaux moyens d'apprentissage. La modalité visuelle leur offre un accès aux informations mieux adapté, mais ils ne disposent pour interagir que des classiques clavier et souris, sans aucune interaction gestuelle.

Nous souhaitons par l'intermédiaire de cette thèse contribuer à l'évolution des connaissances dans le domaine de la reconnaissance et de la compréhension des gestes afin de permettre une avancée à la fois dans le domaine de la communication homme-machine multimodale, mais aussi dans le domaine plus spécifique des didacticiels dédiés à la langue des signes. Nous allons montrer que bien que de gros progrès doivent être réalisés dans le domaine des dispositifs de capture, certains aspects aussi bien des gestes co-verbaux que de la langue des signes peuvent dès maintenant être modélisés.

Le Chapitre 1 présente les caractéristiques des différents types de gestes étudiés dans cette thèse qui sont les gestes de commande, les gestes co-verbaux et les gestes de la langue des signes, ainsi que leur problématique dans le cadre de la communication homme-machine. Dans ce chapitre nous montrons qu'un geste permet de transmettre plusieurs informations simultanément par l'intermédiaire de paramètres qui transportent en parallèle des informations de types différents.

Avant de proposer un système informatique adapté au canal gestuel, il faut étudier plus finement comment sont construits et utilisés ces paramètres. Nous avons choisi d'étudier les gestes de la LSF, car leur structure syntaxique et sémantique est bien définie et leur interprétation ne dépend pas d'une autre modalité. De plus, cette langue est très riche car elle autorise la construction de signes durant le discours en fonction du contexte. Pour ce faire nous avons réalisé et exploité une base de données contenant la description des signes

contenus dans le premier tome du dictionnaire de Bill Moody. Cette étude, présentée au Chapitre 2, a pour rôle d'informer sur les principes de construction et d'utilisation des signes afin de déterminer l'architecture d'un système de reconnaissance et de compréhension dédié au canal gestuel. Elle permet aussi d'aider à la construction de corpus de gestes pour évaluer ce système. Nous montrons dans ce chapitre que deux catégories de signes doivent être distingués : les signes standard, pour lesquels les quatre paramètres sont invariables quelque soit le contexte, et les signes variables, pour lesquels au moins un des paramètres est variable en fonction du contexte.

Le Chapitre 3 présente un système de reconnaissance de gestes de la LSF. Ce système prend en compte les remarques exprimées dans le chapitre précédent, à savoir qu'il existe deux catégories de signes, qui devront être traités différemment. Pour le concevoir, une étude des différents systèmes existants a été effectuée afin de choisir la technique la mieux adaptée à notre objectif, la reconnaissance de phrases de la LSF, c'est-à-dire la reconnaissance de gestes dynamiques enchaînés. La technique de reconnaissance utilisée est basée sur les modèles de Markov cachés qui ont permis l'obtention des taux de reconnaissance très encourageants, sur les deux catégories de signes. Ce chapitre présente en détail les différents outils qu'il a fallu développer afin de mettre au point le système de reconnaissance ainsi que les expérimentations réalisées.

L'objectif du Chapitre 4 est de proposer un système de compréhension de phrases de la LSF. Ce système doit permettre d'interpréter les deux catégories de signes distinguées. Il est connecté au système de reconnaissance décrit dans le Chapitre 3. Il est basé sur la définition de règles spatio-temporelles et la modélisation de la scène de narration qui permet une prise en compte du contexte pour l'interprétation des signes variables. Après un présentation des différents modules qui composent le système, le fonctionnement du prototype qui a été développé est illustré sur un exemple. Différents types d'applications peuvent être développés à partir du système présenté dans ce chapitre.

La dernière partie présente les conclusions et perspectives relatives au travail réalisé dans le cadre de cette thèse.

Une annexe comporte tous les résultats numériques issus de l'exploitation de la base de données conçue durant cette thèse et dont l'analyse est présentée dans le Chapitre 3, ainsi que différentes informations complémentaires.

Chapitre 1

LA COMMUNICATION GESTUELLE

Ce chapitre a pour but de présenter les caractéristiques des différents types de gestes étudiés dans cette thèse qui sont les gestes de commande, les gestes co-verbaux et les gestes de la langue des signes, ainsi que leur problématique dans le cadre de la communication homme-machine.

Il est divisé en cinq sections. La première situe l'étude au sein des différents thèmes de recherche dédiés au canal gestuel. Les trois sections suivantes présentent respectivement les caractéristiques du geste de commande, du geste co-verbal puis du geste de la langue des signes. La dernière section présente les caractéristiques communes à ces trois types de gestes. Elle indique pourquoi nous avons choisi d'étudier plus particulièrement les gestes de la LSF, Langue des Signes Française.

1.1. POSITIONNEMENT DU PROBLEME

Les travaux axés sur le canal gestuel en communication homme-machine peuvent faire appel à des domaines de recherche très différents, selon le type d'interaction étudié. Le but de cette section est de situer notre travail au sein de cette diversité, afin de définir les domaines d'application possibles et les problématiques qui y sont liées.

Après avoir donné une description fonctionnelle du canal gestuel, les différents domaines d'application en communication homme-machine sont présentés. Nous indiquons ensuite les contraintes techniques imposées par ce domaine de recherche. Le dernier paragraphe propose un tableau de synthèse plaçant notre champ de recherche au sein de l'ensemble des domaines possibles.

1.1.1. LES TROIS FONCTIONS DU GESTE HUMAIN

Parmi tous les canaux de communication dont dispose l'être humain, le canal gestuel est sans doute l'un des plus riches. Il permet d'agir sur le monde physique et sert de canal d'information. De plus, ce canal fonctionne dans les deux sens, comme moyen d'émission et de réception d'informations. Si l'on considère la main, on différencie trois fonctions complémentaires et imbriquées, que Claude Cadoz définit de la manière suivante [Cadoz C. 1994] :

- **La fonction épistémique**.

  Dans ce cas, la main joue le rôle d'organe de perception. Par l'intermédiaire du sens *tactilo-proprio-kinesthésique*[1], des informations sur la température, l'état de surface, le poids, la forme ou les mouvements d'un objet peuvent être obtenues.

- **La fonction ergotique**.

  Ici, la main joue le rôle d'organe moteur et agit sur le monde physique pour le transformer. Par l'intermédiaire de l'ensemble de la structure osseuse et des muscles, elle applique à un objet des forces qui vont provoquer une déformation ou un déplacement.

[1] La perception *tactilo-kinesthésique* donne des informations sur la forme, l'orientation, la distance, la grandeur des objets par l'intermédiaire du toucher et de l'utilisation de mouvements exploratoires. La perception *proprioceptive* donne des informations sur le poids, les trajectoires, les mouvements des objets par l'intermédiaire de récepteurs placés dans les articulations et les oreilles [Cadoz C. 1994].

- **La fonction sémiotique**.

  La main joue alors le rôle d'organe d'émission d'information à destination de l'environnement. Elle s'adresse à la perception visuelle d'un ou de plusieurs interlocuteurs.

  Le canal gestuel permet de communiquer une grande diversité d'informations. La puissance d'expression des gestes varie de manière continue. Au bas de l'échelle, se trouvent les gestes faisant partie d'un vocabulaire réduit ne bénéficiant pas d'un enrichissement du message par combinaison des gestes entre eux, tels que les gestes des plongeurs, des grutiers ou des courtiers en bourse. En haut de l'échelle, se trouvent les gestes des langues des signes, qui forment une véritable langue dotée d'une syntaxe et permettent même la création dynamique de signes non standard en fonction des besoins. Entre les deux, se situe le geste co-verbal, qui s'effectue simultanément avec la parole et qui permet d'illustrer ou de compléter le message verbal.

Chacune de ces fonctions prise séparément fait intervenir les deux autres à des degrés variables. Cependant, lorsqu'aucun intermédiaire matériel n'est utilisé et que la fonction sémiotique est prépondérante, les fonctions ergotique et épistémique sont peu exploitées. Elles sont utilisées surtout pour contrôler le mouvement effectué.

Dans le cadre de la communication homme-machine, nous étudierons plus particulièrement cette fonction sémiotique.

1.1.2. LES DOMAINES D'APPLICATION EN CHM

Les fonctionnalités présentées précédemment caractérisent le canal gestuel humain. En communication homme-machine, les domaines d'applications existants à ce jour sont très dépendants de l'évolution des technologies de capture ou de production du geste. On peut distinguer trois grandes catégories d'applications, en fonction du mode d'utilisation du canal gestuel, en entrée-sortie, en sortie et en entrée. Nous citons pour chaque cas des exemples d'applications faisant intervenir le geste dans un espace à trois dimensions :

- **Gestes en entrée-sortie**.
  - En robotique, des organes moteurs permettent de manipuler des objets [LRP 1993]. Dans ce cas, les fonctions épistémique et ergotique de la main humaine sont simulées à l'aide d'un système articulé programmable.

- Dans le domaine de la réalité virtuelle, l'objectif est d'utiliser le geste de la main pour déplacer ou modifier des objets virtuels et pour déclencher une action, par exemple un déplacement de l'utilisateur au sein de la scène. L'objectif est d'exploiter au maximum les possibilités de perception et d'action dont dispose l'utilisateur de manière à ce que son "immersion" soit optimale [Foley 1987].

  Un bon aperçu des différentes problématiques spécifiques à la réalité virtuelle sont présentées dans un numéro spécial de IEEE Computer Graphics & Applications de Janvier 1994 (voir en particulier [Ellis S. R. 1994] pour une définition et un historique de la réalité virtuelle, [Latta J. N. et Oberg D. J. 1994] pour un modèle conceptuel basé sur des études psychologiques, [Encarnação J., Göbel M. et al. 1994] pour une liste des activités européennes dans ce domaine et [Kahaner D. 1994] pour une liste des activités japonaises).

  Dans le but de tester l'apport de la réalité virtuelle en terme d'ergonomie et de convivialité, un atelier de sculpture virtuelle est en cours d'élaboration à l'IRIT (Toulouse) [Torguet P., Rubio F. et al. 1995]. Les gestes sont utilisés pour modeler interactivement des formes virtuelles avec des outils virtuels (marteau, poinçon…).

En entrée-sortie, les trois fonctions se combinent dans ce que Claude Cadoz appelle le geste **instrumental** [Cadoz C. 1994]. Ce type de geste s'applique à un intermédiaire matériel et il y a interaction physique entre l'utilisateur et l'outil utilisé. Il est alors nécessaire de développer des systèmes de capture et de production de gestes assez sophistiqués. Parmi les systèmes permettant de capter des informations en provenance de l'environnement, on peut citer des mécanismes tels que les touches rétroactives, qui permettent de mesurer des mouvements tout en renvoyant une information sur l'effet du mouvement [Gibet S. 1987] (laboratoire ACROE de Grenoble) ou le Dexterous Hand Master du LRP (Laboratoire de Robotique de Paris), qui est un exosquelette attaché sur la main et permet de mesurer les flexions des doigts mais aussi de les modifier [LRP 1993].

- **Gestes en sortie**.

  Dans le domaine de l'animation de personnages virtuels, le but est de produire des représentations graphiques sur écran de personnages réalisant des gestes.

  Certaines recherches [Gibet S. 1992], [Gibet S. et Marteau P.-F. 1994], dont une réalisée au sein du LIMSI [Lebourque T. et Gibet S. 1994], ont pour objectif d'engendrer des gestes, qui sont ensuite affichés à l'écran, à partir d'une description symbolique, dans le cadre de gestes à fonction sémiotique.

[Cassell J., Pelachaud C. et al. 1994] ont implémenté un système qui crée automatiquement des conversations animées entre des personnages virtuels (université de Pennsylvanie - Center for Human Modeling and Simulation). Les conversations sont créées par un planificateur de dialogue qui produit le texte et les intonations des phrases. Les expressions faciales, le mouvement des lèvres, le mouvement de la tête et des bras sont engendrés et synchronisés à partir du texte, de l'intonation et des relations entre le locuteur et son interlocuteur. Les gestes des bras et des mains apportent des informations supplémentaires.

[Lee J. et Kunii T. L. 1993] propose un système qui traduit des phrases de japonais en langue des signes japonaise (Université de Tokyo - Department of Information Science).

- **Gestes en entrée**.

  Les applications les plus couramment rencontrées peuvent être séparées en trois catégories :

  - Les interfaces gestuelles, utilisées pour décrire des formes ou des commandes.
  - Les applications multimodales utilisant les gestes sémiotiques ou ergotiques, le plus souvent associés à la parole.
  - Les applications de reconnaissance de gestes des langues des signes.

  Des travaux illustrant ces domaines d'applications seront présentés plus en détail par la suite.

  Ces trois types d'applications exploitent la fonction sémiotique du geste humain. Ils nécessitent de disposer de systèmes permettant la capture du geste exécuté par l'utilisateur.

1.1.3. LES SYSTEMES DE CAPTURE DE GESTES

Le type de système utilisé pour capter le geste va fortement déterminer les problématiques de traitement des informations mesurées. Pour ce qui est du geste de la main, on distingue deux types de systèmes :

- **Les dispositifs à caméras vidéo**

  Une caméra ne permet de capter que des images en deux dimensions. Si l'on souhaite mesurer les gestes de la main dans l'espace, il est nécessaire de travailler sur une séquence d'images et l'on doit utiliser au minimum deux caméras afin de calculer la troisième dimension. Cela nécessite de développer au préalable des algorithmes d'appariement et de reconstruction. Un problème plus complexe encore

se pose lorsque l'on veut capter les mouvements des doigts, car ces derniers peuvent se cacher les uns les autres (problème d'occultation) et de ce fait, les images peuvent être ambiguës. Il est à ce jour impossible de capter en temps réel les mouvements de la main et des doigts à l'aide de caméras. Par ailleurs, pour capter le mouvement des mains, il est nécessaire de disposer d'une vue assez éloignée de la personne (plan américain) et du coup, même si l'on dispose d'un caméra ayant une bonne résolution, les mains seront représentées avec peu de pixels, donc un résolution faible.

Certains systèmes permettent de simplifier une partie des traitements. Nous avons en particulier testé les deux systèmes suivants utilisant tous deux des marqueurs :

- Le système *Elite*[2] utilise des marqueurs passifs qui sont des pastilles placées sur le corps et réfléchissant une lumière infrarouge émise par des caméras spéciales.
- Le système *Selspot* utilise des marqueurs actifs qui sont des diodes numérotées placées sur le corps et émettant de la lumière infrarouge qui est captée par des caméras infrarouge [Loomis J., Poizner H. et al. 1983], [Poizner H., Klima E. S. et al. 1986].

Ces deux systèmes permettent de capter uniquement les points utiles et de diminuer la quantité d'information à traiter. Cependant, restent les problèmes de disparition de ces points lorsqu'ils sont cachés par rapport aux caméras. Il n'est pas trivial de choisir les bons emplacement pour les marqueurs. De plus, pour le système Elite, l'appariement des points entre les différentes images doit être calculé a posteriori, puisqu'ils ne sont pas différenciés. Précisons que pour le système Selspot, l'utilisateur se trouve connecté à la machine par l'intermédiaire d'un faisceau de fils transmettant l'énergie nécessaire à l'émission de lumière infrarouge par les diodes.

- **Les gants numériques**

Un gant numérique, s'il est associé à un système de capture de l'emplacement et de l'orientation, permet de mesurer le mouvement de la main dans l'espace [Eglowstein H. 1990].

Les problèmes inhérents au domaine de la vision sont ainsi évités. Cependant, d'autres problèmes existent. Il s'agit surtout de problèmes de précision ou de pauvreté des mesures pour certains types de gants, comme cela est précisé dans le Chapitre 3 consacré à la reconnaissance.

---

2 Contact : Société Actisystem - 15 place Grangier - 21000 Dijon

Il est aussi souvent reproché aux gants numériques de contraindre les gestes de l'utilisateur par le fait que celui-ci se trouve "attaché" à l'ordinateur par un fil qui sert à transmettre les données captées. On peut cependant imaginer qu'à long terme il sera sans doute possible de concevoir des systèmes déconnectés de la machine et utilisant un transfert par infrarouge par exemple.

Finalement, le choix se résume de la manière suivante :

- A ce jour, si l'on veut traiter des gestes simples pour lesquels la configuration de la main a peu d'importance, on peut utiliser de simples caméras. Mais le traitement en temps réel ne sera pas toujours possible.
- Si l'on veut pouvoir bénéficier en temps réel de toute la richesse d'information que peut véhiculer un geste de la main, dans l'état actuel de la recherche sur les capteurs de gestes, il sera préférable d'utiliser des gants numériques.

1.1.4. THEME DE RECHERCHE ETUDIE

Ce travail de thèse porte sur les gestes de la main dans l'espace, ayant une fonction sémiotique et captés au moyen d'un gant numérique. Trois types de gestes, représentant différentes étapes dans la puissance d'expression, ont été abordés : nous avons étudié les caractéristiques des gestes de commande dans le cadre d'interfaces gestuelles, des gestes co-verbaux dans le cadre d'applications multimodales, puis des gestes de la Langue des Signes Française (LSF), dans le but de proposer un système de reconnaissance et de compréhension basé sur leurs caractéristiques communes.

Les caractéristiques de ce travail de thèse sont résumées dans le tableau ci-après.

Espace de travail	4 dimensions (espace + temps)
Fonction	sémiotique
Articulateurs	main droite et bras droit
Système de capture	gant numérique et capteur de position/orientation
Types de gestes	gestes de commande gestes co-verbaux gestes de la LSF

Tableau 1.1 : Caractéristiques du travail de thèse.

1.2. LE GESTE DE COMMANDE

Précédemment, nous avons placé les interfaces gestuelles au bas de l'échelle représentant la puissance d'expression du canal gestuel en communication homme-machine. Afin de justifier cette appréciation, nous préciserons d'abord ce que l'on entend par interface gestuelle. Puis une application test réalisée afin d'évaluer un modèle d'interaction gestuelle dans le cadre d'une interface utilisant un langage de commande est présentée.

1.2.1. LES INTERFACES GESTUELLES

Dans ce paragraphe, la terminologie employée dans le domaine des interfaces gestuelles est précisée, puis nous indiquons les caractéristiques de ces dernières en tant que moyen d'interaction entre un utilisateur et une application informatique.

1.2.1.1. Terminologie

Langage de commande et manipulation directe

Le terme *commande* est issu du domaine des interfaces homme-machine. On peut distinguer deux grandes classes d'interfaces suivant le style d'interaction entre l'ordinateur et l'utilisateur : les interfaces à langage de commande et les interfaces à manipulation directe [Delannoy J. F. et Lula J. B. 1990].

Le **langage de commande** peut être un langage artificiel, comme celui des commandes d'un système d'exploitation tel que *UNIX*, ou encore un jeu de commandes par codes, à taper au clavier, utilisé dans les éditeurs de texte tels que *vi*. Il peut être aussi le langage "naturel" ou du moins un sous-ensemble de ce langage. Dans tous les cas, le dialogue est de type *séquentiel* puisqu'il n'est possible d'écrire, de dire ou de lire qu'un mot après l'autre.

La **manipulation directe** est celle qui permet d'agir sur des objets par l'intermédiaire de leur représentation graphique à l'écran [Shneiderman B. 1987]. Les interfaces à manipulation directe sont définies dans un environnement où les commandes et l'affichage font intervenir des objets graphiques. L'interaction est effectuée par la désignation de ces objets à l'aide d'un curseur piloté par une souris. Plusieurs tâches sont accessibles à l'utilisateur (ouvrir un fichier, éjecter une disquette, lancer une application…). Le type du dialogue est alors dit *asynchrone* ou *à événements*.

Mode	Conversationnel	Graphique
Type de dialogue	Séquentiel	Asynchrone
Style d'interaction	Langage de commande	Manipulation directe

Tableau 1.2 : Modes d'interaction.

Classification des périphériques d'entrée

A partir de la manière dont les périphériques d'entrée envoient leurs informations, on distingue deux modes de fonctionnement possibles [Baecker R. M. et Buxton W. A. S. 1987] :

- Le **mode discret** : le périphérique envoie des informations à des instants aléatoires, choisis par l'utilisateur. Ce type de périphérique est dit *à changement d'état discret*.
- Le **mode continu** : le périphérique envoie des informations de façon permanente. Il est dit *à changement d'état continu*.

A partir de ces deux modes, une classification largement répandue consiste à diviser les périphériques en trois catégories :

- Les périphériques fonctionnant en mode discret : on peut citer les claviers, les lecteurs de code à barres, les systèmes de reconnaissance vocale en mots isolés…
- Les périphériques fonctionnant en mode continu : on y trouve les oculomètres (dispositifs permettant de mesurer la direction du regard), les caméras, les gants numériques, les systèmes de reconnaissance vocale en parole continue…
- Les périphériques pouvant fonctionner dans les deux modes : c'est le cas de la souris, du manche à balai, du bouton rotatif ou encore de l'écran tactile.

En pratique, pour les périphériques de la deuxième catégorie, on cherche souvent à se ramener à la troisième catégorie : lorsqu'un périphérique peut fonctionner en mode continu, en général, il peut également fonctionner en mode discret. Il faut pour cela lui adjoindre un module d'analyse, parfois même un module de reconnaissance de formes (gestes, mots, images…). Ce module doit analyser les informations fournies par le périphérique et déclencher un événement dès qu'un signal "utile" est détecté. On peut ainsi obtenir une suite de commandes avec leurs paramètres, que l'on peut interpréter comme les données d'un périphérique à états discrets et intégrer à un modèle classique de gestion des données fournies par l'utilisateur.

Les périphériques fonctionnant en mode continu peuvent traduire une évolution par une succession d'états rapprochés, qui donne une sensation de continuité. Dans le cas du geste, les informations captées par échantillonnage permettent de calculer l'ampleur du geste, ses paramètres d'accélération et ses variations géométriques, avec plus ou moins de précision selon le périphérique utilisé.

Gestes de commande et de manipulation

On appelle **interface gestuelle** tout type d'interface visant à exploiter de façon optimale les informations sur la dynamique gestuelle fournies par la seconde catégorie de périphériques citée précédemment (mode continu). D'où les définitions suivantes :

Les **gestes de commande** constituent un langage gestuel artificiel, utilisé au sein d'une interaction de type conversationnel non graphique.

Les **gestes de manipulation** permettent d'agir sur les objets réels ou virtuels avec un contrôle direct effectué par l'intermédiaire de leur représentation graphique à l'écran, ou, dans le cas ou les fonctions ergotique et épistémique interviennent, par l'intermédiaire du retour tactile ou du retour d'effort.

1.2.1.2. Caractéristiques des interfaces gestuelles

Ce paragraphe résume les points clés de ce type d'interface suivant deux volets : les apports et les problèmes à résoudre.

Apports

Malgré les progrès considérables réalisés quant aux performances des ordinateurs, à leur rapidité, à l'affichage, les interfaces d'entrée semblent n'avoir bénéficié que de peu ou d'aucune amélioration. Or le temps passé par les utilisateurs à entrer des données dans l'ordinateur constitue en fait un goulet d'étranglement. L'amélioration de la rapidité des ordinateurs ne suffira pas à pallier cette difficulté. Il est clair que l'amélioration des systèmes d'entrée entraînera l'amélioration de la productivité des utilisateurs en général [Rubine D. 1991a].

Des progrès ont certes été réalisés et en particulier avec les interfaces de type "*click and drag*", dans lesquelles l'utilisateur manipule directement des objets graphiques sur l'écran à l'aide de la souris. Cette utilisation très intuitive a d'ailleurs provoqué le développement intensif des interfaces à manipulation directe.

Le Macintosh est considéré comme ayant une des meilleures interfaces de type manipulation directe [Foley J., Van Dam A. et al. 1990]. Le "finder" du Macintosh, qui peut être utilisé pour accéder à toutes les applications et à tous les documents Macintosh, est entièrement contrôlé à l'aide de trois techniques d'interaction de base :

- le **clic** : pression rapide sur le bouton de la souris ;
- le "**drag**" : durant un déplacement de la souris, maintien du bouton enfoncé (on gardera le mot anglais *drag*, plus concis, pour représenter cette action)
- le **double-clic** : deux clics simples se suivant dans un court intervalle de temps.

La limitation du nombre de ces techniques est une des raisons de la simplicité d'utilisation du Macintosh. Cependant, elle entraîne aussi une sévère limitation pour bien des applications. Ces dernières devront toujours être construites de manière à ce que leurs opérations s'expriment en termes de ces trois techniques. Dans la pratique, la plupart des applications permettent plus de trois opérations sur un objet. Les opérations autres que celles citées précédemment doivent se faire en plusieurs étapes.

Par exemple, pour déplacer du texte, il faut tout d'abord le sélectionner (à l'aide d'un double-clic ou d'un "drag"), ensuite choisir l'option *Couper* du menu *Édition* (à l'aide d'un "drag") puis sélectionner un emplacement d'insertion (à l'aide d'un clic) et enfin sélectionner l'option *Coller* du menu Édition (avec un "drag"). Le prix à payer, pour la simplicité apportée par le fait de disposer seulement de trois techniques d'interaction, est que certaines opérations sont obligatoirement réalisées à l'aide d'une séquence d'interactions primitives. Le coté intuitif est ici beaucoup moins évident. En effet, certains utilisateurs en viennent à se servir de *raccourcis-clavier*, c'est-à-dire d'un langage de commande, appris au fil du temps (exemples : "□ v" pour copier, "□ c" pour coller… très utilisés pour ce mémoire !). Les opérations complexes peuvent aussi se faire en combinant des interactions souris et clavier (exemple : "shift-clic" pour sélectionner plusieurs fichiers, "□ drag" pour copier du texte sélectionné au préalable).

Un des intérêts des interfaces gestuelles est qu'elles permettent d'éviter cette division des commandes en opérations primitives, en donnant la possibilité de regrouper dans une même opération de base la commande et les paramètres.

A titre d'exemple de ce type d'application, on peut citer GEdit, qui est un prototype d'éditeur graphique 2D permettant de créer et manipuler trois types d'objets simples (cercle, carré, triangle), en utilisant des informations contenues dans certains gestes réalisés avec une

souris [Kurtenbach G. et Buxton B. 1991]. Ces gestes sont sténographiques et de type "correction d'épreuve". Les manipulations réalisées peuvent être l'ajout, la suppression, le déplacement ou la copie. On peut en un geste sélectionner un groupe d'objets et le recopier dans un endroit donné de l'écran.

Dans l'exemple présenté Figure 1.1, le geste consiste à entourer un groupe d'objets (un triangle, un carré, un cercle), aller jusqu'à l'endroit où ils doivent être copiés, puis tracer un "C", pour spécifier la commande, le tout en un seul trait.

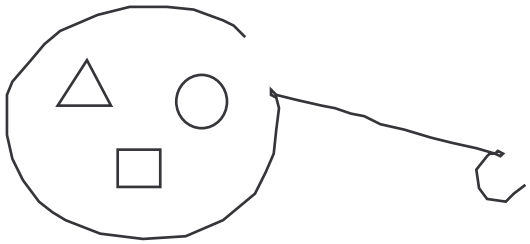


Figure 1.1 : Copie d'un groupe d'objets dans GEdit.

Le geste est **concis** [Buxton W. 1987] car il contient à la fois l'action et le domaine d'action (les paramètres). Il permet de transférer simultanément plusieurs types d'informations. Dans notre exemple, le geste consistant à encercler une partie de texte et à finir le cercle par un C indiquant l'endroit où copier les objets remplace trois composantes linguistiques : le verbe (copier), l'objet direct (les objets graphiques entourés) et le complément circonstanciel de lieu (le lieu de copie).

Les gestes sont **iconiques**. Il existe une relation intuitive entre le geste et ce qu'il représente. Par exemple, lorsque l'on entoure un objet, il est immédiat que le but recherché est de sélectionner l'objet. L'iconicité des gestes permet de mémoriser ces derniers rapidement et même, parfois, de ne pas avoir besoin de les mémoriser.

Il est possible d'**éviter** au moins en partie **la confusion des modes**. En effet, un des moyens utilisés pour éviter de trop longues séquences de clics et de "drags" est de définir des modes dans lesquels des opérations données n'entraîneront pas la même action.

Un exemple est le logiciel MacDraw avec les modes *sélection*, *rectangle*, *cercle*. Ce système est source de nombreuses erreurs, car même si le mode courant est bien affiché, il n'est pas rare que l'on effectue une opération dans un mode, tout en croyant être dans un autre, et que l'on n'obtienne pas le résultat désiré. L'utilisation d'interfaces gestuelles peut

permettre d'éviter certains modes, car le geste comporte intrinsèquement l'information servant à distinguer le mode. Par exemple, dans le cas de l'éditeur de dessin GEdit, le geste tracé suffit à préciser si l'on crée un cercle, un rectangle (Figure 1.2), un triangle, ou s'il s'agit d'une commande [Kurtenbach G. et Buxton B. 1991].

Figure 1.2 : Création d'un rectangle.

Les interfaces gestuelles peuvent **remplacer d'autres modalités** lorsque l'utilisateur n'a pas accès facilement à ces dernières (par exemple, à cause d'un handicap sensoriel ou moteur) ou lorsque celles-ci ne sont pas disponibles. Ainsi, lorsque le fond sonore est trop important, cela interdit l'utilisation d'interfaces vocales. Ces dernières doivent alors être remplacées par une modalité différente, offrant les mêmes fonctionnalités.

Un autre objectif est de se rapprocher de la communication naturelle humaine afin de rendre la **communication** avec les ordinateurs **plus intuitive**. Cette dernière fait en général intervenir plusieurs modalités de communication, telles que la parole, le geste, la direction du regard, la mimique. Le message est alors une résultante de la combinaison des informations issues de ces différentes modalités.

Dans ce but se développent des études sur les interfaces multimodales. Les recherches concernant ce domaine permettent de concevoir des applications utilisant la parole et le geste 2D (au moyen d'une souris, d'un stylo ou d'un écran tactile) pour manipuler des objets graphiques 2D [Bellik Y. 1991], [Faure C. et Julia L. 1992], ou encore un clavier braille dans le cadre d'un outils de traitement de texte pour personnes aveugles [Bellik Y., Pican N. et al. 1994].

Une des perspectives possibles de cette thèse est de doter d'une entrée gestuelle en 3D une application (Mix3D) permettant de manipuler des objets graphiques 3D, utilisant pour l'instant la parole et la souris [Bourdot P., Krus M. et al. 1995].

Les apports des interfaces gestuelles sont indépendants du type du périphérique utilisé pour leur capture [Morrel-Samuels P. 1990].

Problèmes à résoudre

La conception d'interfaces gestuelles passe obligatoirement par la conception d'un système de reconnaissance adéquat. Se posent alors tous les problèmes relatifs à la reconnaissance de formes :

- Lorsque les gestes sont enchaînés, il est nécessaire de les séparer les uns des autres afin d'obtenir une séquence de symboles. On doit trouver un moyen de détecter le début et la fin d'un geste. Il s'agit du processus de **segmentation**.
- Le début d'un geste peut être modifié par la fin du geste précédent, de même que la fin d'un geste peut être modifié par le début du geste suivant. Il s'agit du phénomène de **coarticulation**.

  Le processus de segmentation, faisant intervenir le phénomène de coarticulation, est classique en reconnaissance des formes continues. Il est traité de manière différente, voire opposée, selon que l'on utilise une approche analytique ou une approche globale.

  Dans l'approche analytique, la segmentation est réalisée au préalable et chaque geste isolé des autres est envoyé au module de reconnaissance. On considère que le début et la fin d'un geste sont détectables automatiquement, et qu'il n'est pas nécessaire d'apprendre au préalable toutes les coarticulations possibles.

  Dans l'approche globale, l'ensemble du signal est envoyé au module de reconnaissance, qui procède à une reconnaissance globale de la séquence de gestes, et qui en déduit la segmentation. Il est nécessaire d'apprendre au préalable au système de reconnaissance de nombreuses séquences de gestes enchaînés, afin que celui-ci apprenne les différentes coarticulations possibles.
- La **variabilité** d'exécution d'un même geste suivant la personne, ce que P. Morrel-Samuels nomme le *problème stylistique*, et le contexte, la fatigue ou encore l'humeur de l'utilisateur, qu'il appelle le *problème situationnel*, peuvent rendre la reconnaissance très délicate [Morrel-Samuels P. 1990].

  Le choix d'un algorithme de reconnaissance adéquat et une conception soignée de l'interface à l'aide d'un langage de gestes simples, intuitifs et dont la sémantique est intrinsèquement liée à la forme du geste, peuvent permettre de remédier au moins en partie aux problèmes stylistique et situationnel.

- Comme dans tous les cas où l'on doit réaliser l'interprétation d'un phénomène, des problèmes d'**ambiguïté** peuvent se poser (un même geste peut posséder plusieurs significations possibles), qui doivent être résolus par une prise en compte du contexte.

De plus, des problèmes spécifiques au geste apparaissent :

- Les interfaces gestuelles sont inadaptées à la souris. En effet la souris ne permet de capter qu'une trace 2D de gestes très simples et de ce fait limite fortement la "bande passante" de ce type d'interface (par exemple, la maniabilité limitée de la souris rend difficile la réalisation de gestes de type écriture). D'où la nécessité de disposer de périphériques spécifiques (tablettes graphiques, gants numériques, caméras).
- Il est admis qu'une interface conviviale se doit de présenter autant que possible les actions et les commandes qui sont à la disposition d'un utilisateur dans un état donné. Mais les commandes gestuelles sont souvent difficilement représentables pour l'utilisateur du fait qu'elles sont dynamiques et que leur trace peut s'inscrire dans trois dimensions. Il s'agit du *problème de description*.

  Ce dernier problème reste à ce jour totalement ouvert et n'a pas fait l'objet d'une étude particulière dans le cadre du travail présenté ci-après. On notera toutefois un bon nombre de travaux sur ce thème, dont la plupart sont dédiés à la langue des signes pour laquelle le problème de la notation est capital [Lee J. 1994], [Prillwitz S. et Leven R. 1989]. Une étude comparative des méthodes de transcription de gestes les plus couramment utilisées est donnée dans [Martin-Dupont X. 1995]. L'auteur donne sa préférence au système de notation HamNoSys pour plusieurs raisons. Tout d'abord, les signes idéographiques utilisés sont facilement décodables car ils sont dotés d'une logique interne. De plus, le système est ouvert, puisqu'il permet de transcrire aussi bien les gestes des langues des signes que les gestes co-verbaux. Enfin, il est à la fois précis et économe, aussi bien pour décrire la configuration des différents articulateurs que leur mouvement.

Le choix d'une stratégie de reconnaissance (analytique ou globale) dépend du type d'interaction gestuelle désiré. Dans le cas d'une interface gestuelle pour laquelle le vocabulaire est réduit à une quinzaine de gestes de base, une stratégie de type analytique peut suffire, comme le montre l'étude présentée ci-après.

1.2.2. UN MODELE D'INTERACTION GESTUELLE

Un des apports du geste est qu'il permet de véhiculer plusieurs informations en parallèle. Dans le cadre des interfaces homme-machine, une réelle utilisation du parallélisme du canal gestuel a rarement été étudiée. Elle ne l'a quasiment jamais été pour les gestes 3D captés à l'aide d'un périphérique continu.

C'est le sujet de l'étude suivante, où nous avons cherché à développer une interface gestuelle possédant les qualités définies précédemment.

1.2.2.1. Objectif

L'objectif de cette étude est de proposer une définition des interactions gestuelles ainsi que des algorithmes de reconnaissance dans un domaine d'application restreint : les interfaces gestuelles [Braffort A., Baudel T. et al. 1992], [Baudel T. et Braffort A. 1993]. Le système de capture de gestes utilisé est un gant numérique.

Ce travail de DEA réalisé au LIMSI [Braffort A. 1992] a fait l'objet d'un document vidéo présentés lors de différents séminaires [Baudel T. et Braffort A. 1992].

1.2.2.2. Définition des interactions gestuelles

Comme nous l'avons vu précédemment, le geste peut véhiculer simultanément des informations relatives à une action et aux paramètres permettant de préciser l'étendue de l'action. Le premier type d'information peut être assimilé à un verbe, tandis que le second correspond soit au sujet, soit au complément associé à un verbe.

Il est clair que ces deux types d'informations n'ont pas de sens s'ils sont pris isolément. Une interaction sera toujours constituée d'au moins une information de chaque type. Il va être nécessaire de définir une grammaire permettant d'associer ces deux types d'informations. Comme elles peuvent être simultanées, cette grammaire portera à la fois sur la succession des gestes et sur leur structure parallèle.

L'idée développée dans cette étude a consisté à affecter à des parties du geste distinctes les différents types d'informations. Le mouvement est utilisé pour communiquer la commande, tandis que la configuration (forme de la main) permet de transmettre les paramètres de la commande. Ce choix n'est pas arbitraire car il correspond aux structures sémantiques du geste humain, comme nous le verrons par la suite.

Plus précisément, le geste a été structuré en trois parties :

- la configuration statique de début, pouvant transmettre une information de type paramètre ;
- le mouvement, pouvant transmettre une information de type commande ;
- la configuration statique de fin, pouvant transmettre une information de type paramètre.

On bénéficie ainsi de la simultanéité d'information et de plus, ce découpage dans le temps permet de réaliser facilement la segmentation des gestes lorsqu'ils sont enchaînés. Il suffit d'avoir une liste des configurations de début et de fin valides pour l'application pour pouvoir déterminer le début et la fin d'un geste dynamique.

1.2.2.3. Algorithme de reconnaissance

Le module de reconnaissance développé prend en compte cette structuration du geste en trois parties. Nous en donnons une rapide description. Il fonctionne en quatre étapes :

- Une comparaison de la configuration courante par rapport à la liste des configurations de début valides est effectuée en continu, jusqu'à ce qu'une configuration de début CD soit détectée à l'instant T1.
- Une comparaison de la configuration courante par rapport à la liste des configurations de fin valides est effectuée en continu, jusqu'à ce qu'une configuration de fin CF soit détectée à l'instant T2. Entre T1 et T2, les valeurs issues du gant numériques sont mémorisées.
- Cette suite de valeurs est envoyée au module de classification, afin d'être reconnue. On en déduit une action A.

  Ce module de classification est issu d'un module utilisé pour la reconnaissance de gestes 2D [Rubine D. 1991a]. Il a été adapté pour les gestes 3D. Son principe repose sur le calcul incrémental des caractéristiques géométriques du geste (longueur totale du geste, angle de départ, diagonale de la boîte englobante…). Il sera détaillé au Chapitre 3 sur la reconnaissance.
- A l'aide de règles simples permettant d'interpréter l'ensemble CD, CF et A, on obtient la commande ainsi que ses paramètres.

Cet algorithme permet de reconnaître des gestes 3D dynamiques enchaînés.

1.2.2.4. Application test

Afin d'évaluer le modèle, une application test a été développée, nommée "Charade". L'application consiste à commander, à l'aide de gestes de la main, la navigation au sein d'un ensemble de transparents. Ces transparents sont des pages d'écran de l'ordinateur. Une platine de rétroprojection permet de projeter l'écran de l'ordinateur sur un écran mural. Le gant numérique est relié à l'ordinateur, qui analyse en continu les données gestuelles. En fonction des gestes effectués, un nouveau transparent est affiché, ou encore une zone de l'écran est sélectionnée. Ce type d'outil peut être utilisé lors d'une intervention en public (conférence, cours, démonstration…), comme l'illustre la Figure 1.3.

Le fonctionnement du gant numérique est détaillé dans le Chapitre 3.

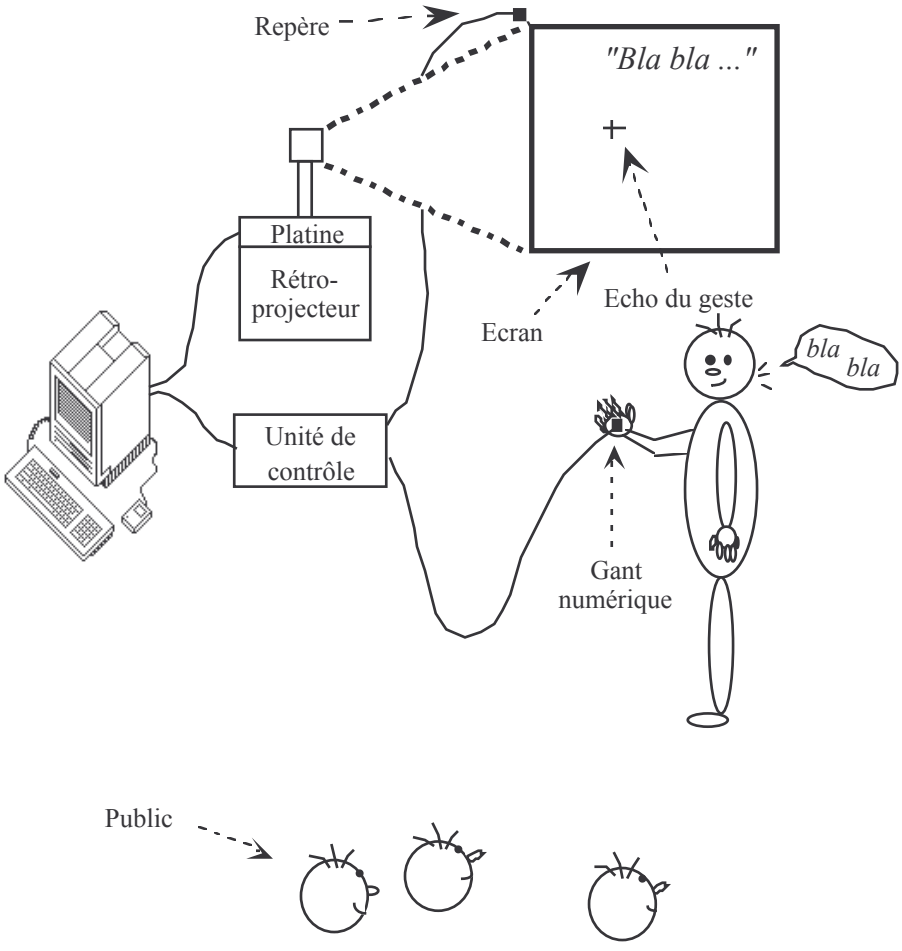


Figure 1.3 : Application test Charade.

1.2.2.5. Vocabulaire gestuel

Le vocabulaire est composé d'une dizaine de gestes simples tels que "*Passer au transparent suivant*", "*… au précédent*". De manière à éviter à l'utilisateur d'avoir à apprendre un code gestuel artificiel, une métaphore assez intuitive a été utilisée, dans le cadre d'un parcours au sein d'une séquence de pages : la manipulation d'un livre. Ainsi, les transparents son identifiés aux pages d'un livre. Passer au transparent suivant correspondra à un mouvement en arc de cercle de la droite vers la gauche, tandis que revenir au transparent précédent sera le même mouvement, mais effectué de gauche à droite. En combinant le geste "*Passer au suivant*" et la configuration représentant le chiffre "*trois*", on peut alors spécifier une commande et son paramètre : "*Passer au troisième transparent*" (Figure 1.4).

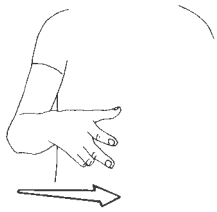


Figure 1.4 : Exemple de geste de commande.

Le vocabulaire comporte aussi des gestes permettant de manipuler des objets graphiques de type boutons, comme l'illustre la Figure 1.5.

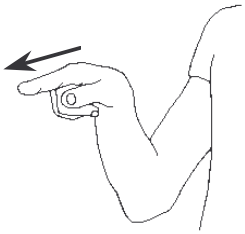


Figure 1.5 : Exemple de geste de manipulation.

Ce type de geste correspond à des gestes de manipulation, tandis que des gestes tels que celui illustré Figure 1.4 correspondent à des gestes de commande, puisqu'ils ne s'appliquent pas à un objet graphique.

Pour cette interface, cohabitent des gestes de commande et des gestes de manipulation. La plupart des actions peuvent être exécutées soit par un geste de commande, soit par un geste de manipulation.

1.2.2.6. Évaluation

Lors des tests en grandeur réelle réalisés afin d'évaluer l'interface, il a été constaté que les gestes qui semblaient les plus intuitifs était les gestes de commande et non pas les gestes de manipulation, comme on pouvait s'y attendre.

Cela peut s'expliquer par le fait que les objets graphiques de l'interface sont des représentations dans un monde à deux dimensions, comme par exemple des boutons pour lesquels l'interaction la plus intuitive est un geste de pression de ce bouton. Les gestes de manipulation proposés simulent bien cette action de pression, mais deux problèmes sont apparus :

- Aucun retour tactile n'étant disponible, l'aspect intuitif du geste s'en trouve amoindrit.
- L'utilisateur étant assez loin de l'écran mural, il faut "viser" le bouton afin d'être sûr de manipuler le bon bouton.

Pour cette application, les gestes de commandes permettent à l'utilisateur d'obtenir une certaine autonomie par rapport à l'application et à l'ordinateur, ce qui n'est pas le cas avec les gestes de manipulation.

Le fait d'utiliser des gestes de la main pour changer de transparent plutôt que de manipuler une souris ou un clavier permet à l'utilisateur de ne pas avoir à manipuler l'ordinateur, qui devient presque transparent pour lui. Il peut alors se concentrer exclusivement à sa tâche, la présentation orale à un public.

Il est nécessaire de connaître le contexte dans lequel est émis un geste, afin de déterminer les intentions de l'utilisateur. La technique[3] utilisée pour distinguer les gestes intentionnels des gestes involontaires était insuffisante car certains gestes ont été reconnus par le système comme étant des commandes, alors qu'il s'agissait de gestes co-verbaux émis en direction du public. Cela provoquait des changements de transparents non désirés (et l'étonnement du public !).

---

[3] Cette technique consiste à n'autoriser la reconnaissance du geste que si un curseur indiquant la direction de la main par rapport à un repère fixe apparait dans la zone de l'écran. Lorsque l'utilisateur est tourné vers le public, il n'a aucun contrôle sur la présence ou l'absence de ce curseur sur l'écran.

Le système de reconnaissance utilisé est assez simple à mettre en oeuvre (une interface graphique permet de définir facilement les configurations valides ainsi que les classes de gestes dynamiques). Les algorithmes utilisés maintiennent une dépendance entre les paramètres de configuration et de mouvement, tant du point de vue de la reconnaissance que du point de vue de la segmentation. Il s'est avéré que cela conduisait à contraindre l'ampleur des gestes et le choix des configurations. De ce fait, les commandes gestuelles utilisées ne sont pas les plus intuitives pour l'utilisateur.

1.2.3. BILAN

Les interfaces gestuelles permettent l'utilisation de gestes de commande et de gestes de manipulation. Comme ces derniers, les gestes de commande permettent de véhiculer plusieurs types d'informations qui se combinent. Ce type d'interface est bien plus "puissant" que les interfaces à langages de commandes classiques, qui sont limitées à une combinaison séquentielle d'informations.

L'utilisation de la simultanéité d'informations permet une nette simplification dans l'utilisation de l'interface gestuelle car le nombre de gestes qui composent le vocabulaire gestuel est sensiblement réduit et facilement mémorisable.

La simplicité de la mise en oeuvre du module de reconnaissance va de paire avec la simplicité et la petite taille du vocabulaire gestuel. Un système de reconnaissance tel que celui utilisé précédemment sera insuffisant dans le cadre d'une application nécessitant un vocabulaire de grande taille. Mais si l'on perfectionne les algorithmes en tenant compte des observations relevées lors de l'évaluation, ce système peut suffire pour des applications ciblées et dont les interactions de base sont bien délimitées.

1.3. LE GESTE CO-VERBAL

Dans le cadre des interfaces gestuelles, nous avons vu qu'un des rôles du canal gestuel est d'apporter plus de naturel dans la communication homme-machine, de manière à la rendre plus efficace. Pour cela, cette communication doit se rapprocher le plus possible de la communication humaine. Cette dernière est par essence multimodale. Le geste et la parole se combinent pour créer des messages à la fois plus complets du point de vue du contenu et plus simples du point de vue de la réalisation.

Il est intéressant d'étudier le geste co-verbal, d'en connaître la structure et le fonctionnement, afin d'être à même de concevoir des interfaces multimodales performantes.

Après avoir présenté une classification des gestes co-verbaux basée sur leur valeur sémantique au sein de la communication, nous présentons une étude portant sur une interface multimodale utilisant la parole et le geste pour communiquer des informations de type spatial. Une étude des gestes exprimant le temps est ensuite exposée. Les caractéristiques principales des gestes co-verbaux sont enfin résumées dans un paragraphe de synthèse.

1.3.1. FONCTIONS DES GESTES CO-VERBAUX

Les chercheurs travaillant sur les gestes co-verbaux ont proposé différentes classifications de ces gestes. Il ne semble pas exister d'unanimité en la matière. Les classifications les plus répandues sont présentées.

1.3.1.1. Gestes symboliques et gestes illustrateurs

Généralement, au moins deux classes sont différenciées, selon que le geste possède une sémantique "autonome" par rapport au message verbal ou pas [Ekman P. et Friesen W. V. 1972], [Mc Neill D. 1992].

- Les gestes **symboliques** ou **emblématiques** sont indépendants du canal verbal. Ils peuvent accompagner ou remplacer un mot ou un groupe de mots. Ces gestes sont propres à des communautés sociolinguistiques.

  Par exemple, le geste de salut est un emblème. On peut prononcer "Au revoir" tout en exécutant le geste, mais dans ce cas, les deux messages sont sémantiquement redondants.

- Les gestes **illustrateurs** ne sont pas indépendants du canal verbal. Le sens complet du message est obtenu en combinant les contenus du message oral et du message gestuel. Par exemple, la phrase "Je veux celle-là !" n'est interprétable que si elle est accompagnée d'un geste désignant par exemple une bouteille spécifique (disons … un Saint Émilion 85) parmi un ensemble de bouteilles.

Parmi les illustrateurs, une sous-classification est possible. Là encore, les avis diffèrent. David Mc Neill [Mc Neill D. 1992] distingue en particulier les sous-classes de gestes suivantes :

- Les gestes **déictiques**. Ce sont des mouvements de pointage, qui sont en général exécutés avec l'index tendu et les autres doigts pliés, mais parfois aussi avec d'autres parties du corps (tête, nez, menton …) ou par l'intermédiaire d'artefacts (règle, stylo …). Ils désignent un objet qui est simultanément référencé dans le discours, comme la bouteille dans l'exemple précédent.
- Les gestes **iconiques**. Ce sont les gestes qui représentent un objet, une action ou un événement concrets, eux-mêmes simultanément référencés dans le discours oral. Ils indiquent la forme, la taille, ou d'autres caractéristiques propres. Par exemple, la phrase "J'ai péché un poisson gros comme ça !" sera accompagnée d'un écartement des deux mains indiquant la taille (plus ou moins exacte !) du poisson.
- Les gestes **métaphoriques**. Ici, les images présentées illustrent un concept abstrait, comme par exemple dans "Il veut s'emparer de nos idées !" accompagné d'un mouvement de la main, mimant l'action de saisir un objet.

  Notons que cette définition est remise en cause par M. de Fornel [De Fornel M. 1993], qui considère qu'elle n'est pas correcte puisque, dans le type d'exemple cité précédemment, ces gestes se contentent de reprendre fidèlement l'expression à laquelle ils sont affiliés par l'intermédiaire d'une traduction visuelle de l'entrée lexicale. Pour lui, les gestes métaphoriques sont définis par le fait qu'il doit y avoir "une incongruité sémantique entre la schématisation conventionnelle du geste et celle de l'expression verbale affiliée". Par exemple, le geste des deux mains qui esquissent les contours d'une sphère, associé à la phrase "le musée était fermé", est un geste métaphorique car il apporte à la phrase une notion de clôture, d'univers clos, coupé du monde.
- Les gestes de **battements** (ou bâtons [Ekman P. et Friesen W. V. 1972]). Ils marquent le rythme du discours. Leur forme est indépendante du contenu sémantique du discours.

Si l'on se rapporte aux interfaces gestuelles étudiées précédemment, on peut considérer que dans l'application test présentée, les gestes de commande correspondent à la catégorie emblème, tandis que les gestes de manipulation correspondent à la catégorie illustrateur. On a vu pour l'application test que les gestes les plus naturellement utilisés étaient les gestes de commande. Cela n'est cependant pas une règle générale. De nombreuses applications nécessitent de manipuler des objets graphiques ou textuels. Les gestes les plus étudiés pour ce type d'application sont les gestes déictiques. Mais des études sur les gestes iconiques commencent à apparaître.

1.3.1.2. Les gestes déictiques

Les gestes déictiques co-verbaux sont en général très simples dans leur réalisation. La configuration de la main est le plus souvent un index tendu avec les autres doigts pliés, parfois une main plate. Le mouvement est en général un petit geste de pointage, dont la trace dans l'espace est une petite droite en direction de l'objet pointé. Parfois même, la main est immobile.

Jusqu'à très récemment, les interfaces multimodales utilisant le canal gestuel se sont focalisées sur le geste déictique. Il n'est pas étonnant que ce type de gestes ait été le premier à être étudié, car ils sont d'utilisation très courante pour une quantité de tâches, telles que l'édition de texte ou la création de dessins et, d'une manière générale, pour toute application utilisant des interactions de type manipulation directe, dans un espace à deux dimensions, mais aussi à trois dimensions.

Ainsi, une des premières études, menée au MIT, "Put-that-there", permet à un utilisateur de créer, nommer, modifier et déplacer des objets sur un écran à l'aide de la parole et de gestes de désignation [Bolt R. A. 1980], [Bolt R. A. 1987]. Plus tard, de nouveaux types d'informations déictiques ont été ajoutées, telles que la direction du regard [Thorisson K. R., Koons D. B. et al. 1992]. Certaines études utilisent le gant comme moyen de détecter des configuration de la main de type désignation (index tendu) [Braffort A. 1992], [Nogier J.-F. 1993].

1.3.1.3. Les gestes iconiques

Classification

Les gestes iconiques sont généralement subdivisés en trois sous-groupes dont les noms varient selon les auteurs. Nous avons repris la terminologie de [Sparrel C. J. 1993] :

- Les gestes **spatiographiques** sont utilisés pour représenter les relations spatiales entre les référents. Ceci est réalisé par l'intermédiaire des emplacements relatifs des gestes, dans un espace de référence créé par l'orateur. Par exemple, les gestes accompagnant la phrase "Le verre est [geste] là et la bouteille est [geste] là." sont spatiographiques. Ces objets sont placés dans la scène par rapport à la position du locuteur.
- Les gestes **pictographiques** sont utilisés pour montrer la forme d'un objet. Ceci est réalisé par l'intermédiaire de la configuration de la main, éventuellement accompagnée d'un mouvement. Par exemple le geste accompagnant la phrase "Cette bouteille a une drôle de [geste] forme." est pictographique.
- Les gestes **kinégraphiques** sont utilisés pour représenter certains types de mouvements, à l'aide d'un déplacement des mains. Par exemple, le geste accompagnant la phrase "Après le huitième verre, il titubait [geste] !" est kinégraphique.

Cette classification en trois sous-groupes a le défaut de ne pas différencier deux types de gestes, ceux qui donnent des informations sur la forme de l'objet et ceux qui donnent des informations sur l'orientation de l'objet. Cela nous a amené à ajouter une quatrième classe :

- Les gestes **orientographiques**[4], qui sont utilisés pour préciser l'orientation d'un objet dans la scène. Souvent, on indique l'orientation de l'objet par rapport à la verticale. Ceci est réalisé par l'orientation des mains. Par exemple, le geste accompagnant la phrase "Attention ! Tu penches [geste] ton verre !" est orientographique.

Paramètres du geste

A partir de cette classification, on constate que le geste de la main est décomposable en quatre parties distinctes, qui véhiculent quatre types d'information différents :

- La **configuration** de la main, sa forme, donne des informations de type pictographique.
- Le **mouvement** de la main donne des informations kinégraphiques, sur le mouvement des objets, mais aussi pictographique, sur la forme des objets.

4 Nous avons choisi ce terme parce que l'information est donnée par l'orientation de la main.

- L'**orientation** de la main donne des informations orientographiques, sur l'orientation de l'objet dans la scène.
- L'**emplacement** de la main donne des informations spatiographiques, sur l'emplacement de l'objet dans la scène.

Ces quatre sources d'information seront nommées dans la suite du mémoire les quatre **paramètres** du geste.

Applications

Les interfaces multimodales utilisant le geste iconique sont encore très rares, d'une part parce que les applications concernées sont moins nombreuses que celles utilisant le geste déictique, mais aussi parce que leur modélisation est nettement plus complexe.

Une des rares études existantes ayant atteint un stade opérationnel est un démonstrateur nommé VECIG (Virtual Environment Coverbal Iconic Gesture) réalisé au MIT [Sparrel C. J. 1993]. L'application consiste à positionner des objets dans une scène 3D virtuelle. Les gestes servent à indiquer la localisation, l'orientation ou le mouvement d'un ou de plusieurs objets, en combinaison avec la parole.

1.3.2. INFORMATIONS SPATIALES MULTIMODALES

Dans le cadre d'interfaces multimodales, on peut s'intéresser aux emblèmes car ils confirment un message verbal. Mais ce sont surtout les illustrateurs qu'il est important d'étudier, tout du moins les gestes déictiques et iconiques, car ils sont indispensables à la compréhension du message.

Ceci est particulièrement vrai lorsque l'on cherche à communiquer des informations de type spatial. Dans le cadre d'une étude prospective, nous avons analysé comment la coopération entre le message vocal et le message gestuel peut permettre de résoudre les ambiguïtés présentes dans l'un ou l'autre des canaux de communication. L'association des deux modalités doit permettre de concevoir des interfaces multimodales proposant à l'utilisateur des interactions puissantes tout en restant naturelles.

1.3.2.1. Types d'information

Le schéma ci-dessous (Figure 1.6) rappelle l'ensemble de toutes les classifications vues précédemment.

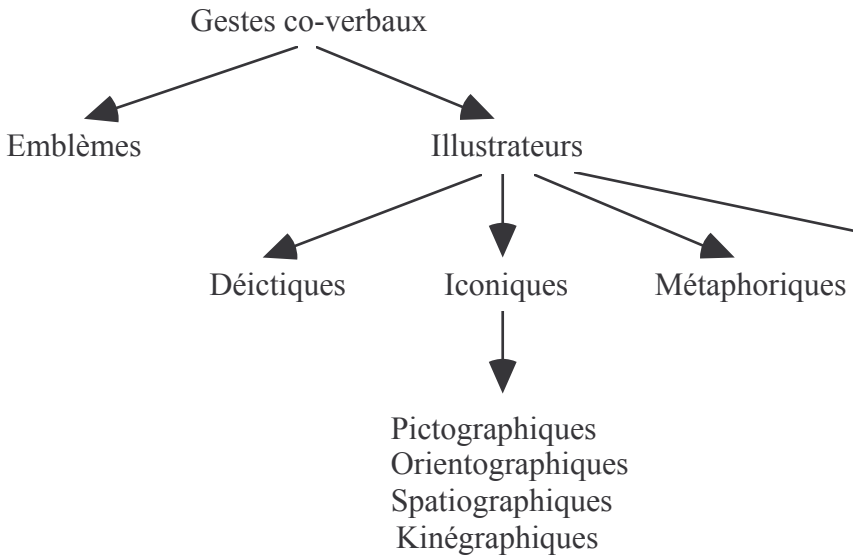


Figure 1.6 : Classification des gestes co-verbaux.

En langage naturel, quatre types de prépositions spatiales peuvent être distingués [Briffault X. 1992]. Ce sont celles qui véhiculent :

- Les informations de **localisation** d'un objet par rapport à un repère donné. Cette catégorie se subdivise en quatre sous-classes, qui sont :
  - Les informations **topologiques** permettant d'exprimer des relations telles que le contact, l'inclusion…
  - Les informations **projectives** sont fournies par la projection de l'objet suivant un repère donné (à gauche, à droite…)
  - Les informations **géométriques** permettant de situer un objet par rapport à d'autres en fonction de la configuration géométrique qu'ils forment (entre, aligné …)
  - Les informations de **distance** d'un objet par rapport à d'autres (près, loin, à 3 kilomètres …).
- Les informations de type **morphologique** indique la forme ou la taille d'un objet (gros, arrondi …).
- Les informations de **position** (dite intrinsèque) d'un objet permettent de connaître l'orientation des axes d'un objet par rapport à un repère donné. Souvent, l'axe de référence est l'axe vertical défini par la gravité (allongé, penché …).

D'autres lexèmes sont utilisés pour communiquer des informations spatiales, en particulier les verbes de mouvement, certains substantifs ou adjectifs.

Dans le cadre d'une communication multimodale, l'utilisation des gestes permet de simplifier le message verbal. Nous donnons ci-dessous quelques exemples.

Pour spécifier un objet parmi un ensemble d'objets, on peut utiliser un geste déictique, comme l'illustre la Figure 1.7.

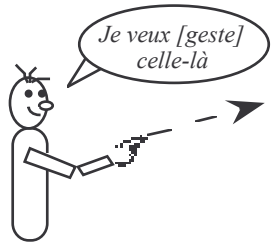


Figure 1.7 : Un exemple de désignation.

Pour indiquer la localisation de plusieurs objets dans une scène, on utilise des gestes spatiographiques, qui permettent de résoudre les anaphores déictiques du message verbal, comme l'illustre la Figure 1.8.

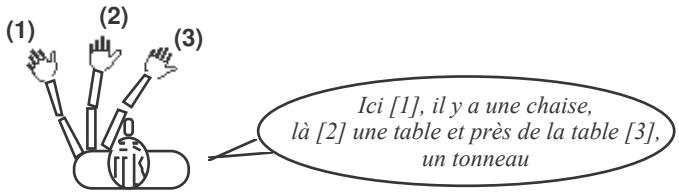


Figure 1.8 : Un exemple de localisation.

Pour indiquer la taille ou la forme d'un objet, on utilise des gestes pictographiques, comme l'illustre la Figure 1.9.

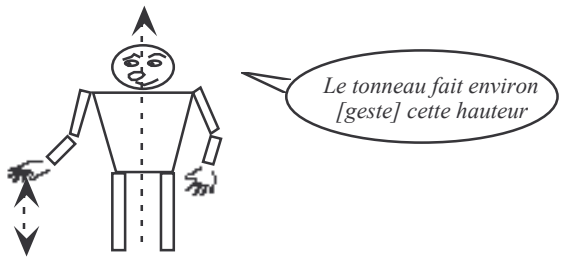


Figure 1.9 : Un exemple d'information morphologique.

Pour indiquer l'orientation d'un objet, on utilise des gestes orientographiques, comme l'illustre la Figure 1.10.

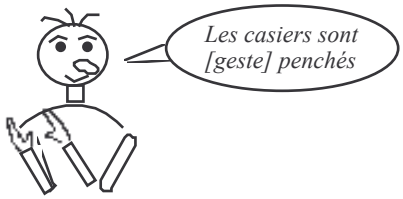


Figure 1.10 : Un exemple d'information de position.

1.3.2.2. Ambiguïté du geste

Au vu de ces exemples, on constate que certains gestes ne véhiculent pas le même type d'information, alors qu'ils sont identiques. Par exemple, dans la Figure 1.11, deux gestes successifs de la main droite sont représentés. Le premier geste [1] véhicule une information de localisation, tandis que le second [2] transmet une information morphologique. Dans le premier cas, ce sont les coordonnées **(x, y)** de la main par rapport à un repère donné qui sont pertinentes, tandis que dans le second cas, c'est la distance **z** entre le sol et la main qui est pertinente.

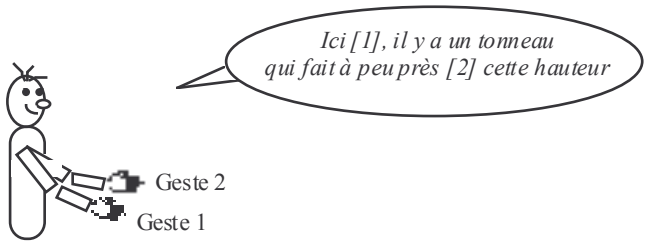


Figure 1.11 : Un exemple de gestes ambigus.

1.3.2.3. Simultanéité d'information

Nous avons constaté par ailleurs qu'un geste peut transmettre plusieurs types d'informations spatiales simultanément, ceci par l'intermédiaire de spécialisation de parties du geste de la main, qui ont été nommées *paramètres* précédemment. Ces quatre paramètres ont la propriété de pouvoir véhiculer quatre informations de type différent simultanément.

Par exemple, le geste présenté Figure 1.10 comporte une information de position, véhiculée par l'orientation des mains qui indique que les casiers sont penchés, mais aussi une information de type géométrique puisque les deux mains parallèles indiquent que les casiers

sont parallèles, et aussi un information de type localisation, donnée par l'emplacement des mains. Si l'on avait ajouté à ce geste statique un mouvement, on aurait pu indiquer par exemple, que les casiers étaient en train de tomber. Remarquons de plus que la forme des casiers est rappelée par la configuration plate des mains.

1.3.2.4. Bilan

Le canal gestuel est particulièrement adapté pour transmettre des informations de type spatial, ce qui n'est pas spécialement le cas du canal verbal. Les gestes co-verbaux sont capables de véhiculer plusieurs informations simultanément, au contraire des mots du langage naturel, de par la structure linéaire de ce dernier.

Par contre, toujours dans le domaine des informations de type spatial, les gestes peuvent être ambigus, puisque leur interprétation dépend du contenu du message verbal. Si l'on veut concevoir une interface multimodale utilisant le langage naturel et les gestes co-verbaux naturels, l'interprétation du message gestuel sera dépendant de l'interprétation du message verbal. Par ailleurs, certains mots du langage naturel sont ambigus et le geste peut permettre de lever ces ambiguïtés. Le module chargé de l'interprétation doit gérer la résolution des ambiguïtés possibles des deux modalités.

Nous montrons ci-dessous les différentes étapes amenant à l'interprétation des messages verbaux et gestuels, dans le cadre d'une application où l'on manipule des objets dans une scène virtuelle. Il est supposé que l'on dispose de systèmes de reconnaissance de mots et de gestes isolés. Dans le but de spécifier un endroit de la scène, l'utilisateur prononce le mot "LA" accompagné d'un geste "main plate".

- Étape de reconnaissance :

  La syllabe "LA" est reconnue.

  Le geste "main plate vers l'avant, paume vers le bas, statique" est reconnu.
- Étape d'interprétation :

  Le mot "LA" peut être :

  un article défini ("la"),

  un adverbe de lieu ("là").

  Le geste peut être :

  un déictique,

  une localisation.

Il en ressort que le seul appariement possible est l'adverbe de lieu avec le geste déictique.

- Étape de mise à jour de la scène :

  On déduit que les valeurs pertinentes sont les valeurs (x, y) dans la scène virtuelle.

Au vu de cet exemple, on voit qu'il est nécessaire de disposer des modules suivants :

Reconnaissance :

1. Un système de reconnaissance dédié au message gestuel et un autre dédié au message verbal.

   Ces systèmes fournissent des représentations symboliques des messages.

   Notons au passage que ces messages ne sont que rarement simultanés. Souvent, mais pas systématiquement, le geste précède la parole [Kendon A. 1980], [Mc Neill D. 1992].

Interprétation :

2. Un système d'analyse syntaxique dédié au message verbal et un autre dédié au message gestuel.

   Ces deux modules fournissent pour chaque message une liste des possibilités. Il faut que pour chaque geste du vocabulaire utilisé dans l'application, tous les types d'informations spatiales qu'il est à même de véhiculer soient définis au préalable.

3. Un module chargé de faire l'intersection entre les deux ensembles créés précédemment.
4. Un module chargé de représenter la scène virtuelle, qui pourra être interrogée ou modifiée.

Cette étude a été menée dans le but d'aider à la réalisation d'une interface multimodale dans le cadre d'une application gérant des informations spatiales [Briffault X. et Braffort A. 1993a], [Briffault X. et Braffort A. 1993b]. Tous les problèmes liés à la multimodalité ne sont pas abordés ici, en particulier les problèmes de gestion des informations temporelles, problèmes cruciaux en multimodalité. De plus, le choix d'une architecture spécifique reste un problème ouvert. Simplement, nous apportons ici quelques remarques dont le concepteur de l'architecture devra tenir compte afin de permettre une utilisation optimale de l'interaction entre les modes gestuel et verbal.

1.3.3. INFORMATIONS TEMPORELLES

Au sein du LIMSI, un groupe de travail rassemble des chercheurs et étudiants d'organismes et d'horizons différents (linguistes, informaticiens, enseignants de langue des signes…) autour d'un thème commun : "Le geste de communication" (voir Annexe 1.1).

Dans le cadre de ce groupe de travail une étude sur les gestes exprimant des informations temporelles a été menée. Deux cassettes vidéo ont été enregistrées sur ce thème, une sur les gestes co-verbaux [Calbris G. 1993] et une autre sur les gestes de la LSF [Cuxac C. 1993b].

A partir de ces deux documents, des discussions menées au sein de ce groupe de travail et de différentes publications sur ce thème [Friedman L. A. 1975], [Moody B. 1983], [Calbris G. 1985], [Calbris G. et Montredon J. 1986], les caractéristiques communes aux deux types de gestes dans le domaine de la communication d'informations temporelles ont été étudiées :

- Localisation par rapport au moment actuel.

  L'emplacement du geste permet d'indiquer si l'action se passe dans le passé, le présent ou le futur. Cet emplacement n'est pas quelconque. Il se situe sur une ligne horizontale placée sur le coté du locuteur ou du signeur[5] (Figure 1.12).

  - Le passé est situé vers l'arrière du locuteur. Le passé récent est situé juste derrière le locuteur, tandis que le passé lointain se situe plus loin vers l'arrière.
  - Le présent est situé vers le bas, aux pieds du locuteur.
  - Le futur est situé vers l'avant du locuteur. Le futur récent est situé juste devant le locuteur, tandis que le futur lointain se situe plus loin vers l'avant.

[5] La qualité de locuteur (langue orale) ou de signeur (langue des signes) n'intervient que peu dans la suite de ce paragraphe. Nous avons regroupé les deux termes sous celui de "locuteur" au sens générique, celui qui émet un message, l'interlocuteur étant celui qui le reçoit.

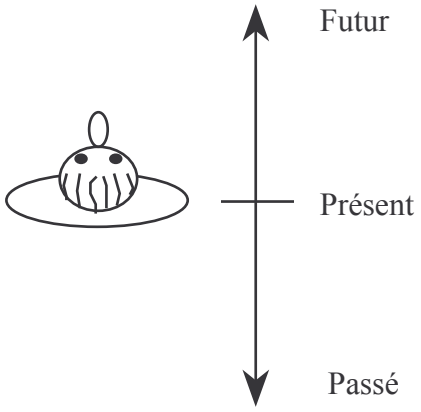


Figure 1.12 : Première ligne de temps (vue du dessus).

- Localisation par rapport à un instant T.

  Dans ce cas, une ligne horizontale frontale est utilisée. Les emplacements relatifs de la main le long de cette ligne permettent de classer chronologiquement les événements les uns par rapport aux autres (Figure 1.13).

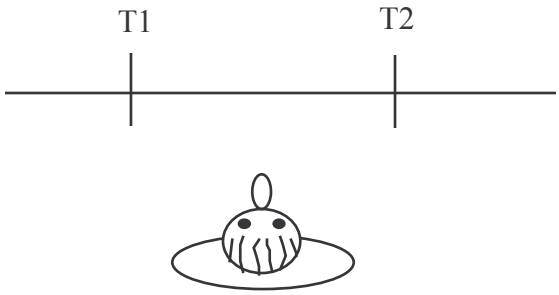


Figure 1.13 : Deuxième ligne de temps (vue du dessus).

  Cet axe n'est pas orienté mais souvent, T1 est situé sur la gauche quand T1 < T2.

- Primitives de mouvement.

  Les primitives géométriques que tracent les gestes dans l'espace sont en nombre restreint. On observe des droites, des arcs, des cercles sur place et des cercles avec translation.

- Dynamique du mouvement.

  La dynamique du mouvement véhicule des informations sur la manière dont se déroule une action dans le temps, telles que la répétition, la succession, l'arrêt brutal, la continuité.

Si l'on reprend la classification utilisée pour les informations de type spatial, on remarque les points suivants :

- Un geste déictique permet de préciser la notion de passé/présent/futur puisqu'il permet de "pointer" vers différentes zones de l'espace.
- Un geste spatiographique permet de positionner des événements dans le temps les uns par rapport aux autres puisqu'il permet d'indiquer des relations spatiales.
- Un geste pictographique permet d'indiquer des durées car il permet d'exprimer des tailles.
- Un geste kinégraphique permet de décrire le déroulement dans le temps par l'intermédiaire de la dynamique du mouvement.

Ici encore, les différents paramètres de la main se spécialisent pour véhiculer des informations différentes, cette fois-ci temporelles.

1.3.4. CONCLUSION

En langage naturel, les mêmes prépositions, verbes et adverbes peuvent être employés pour exprimer indifféremment des informations spatiales et temporelles. L'expression gestuelle du temps confirme la conception d'un même continuum espace-temps. L'espace parcouru se confond avec le temps écoulé.

Pour ces deux domaines, spatial et temporel, on observe que les paramètres de la main (configuration, mouvement, orientation et emplacement) se spécialisent pour véhiculer des informations de types différents. Ceci permet une grande efficacité du message gestuel car plusieurs informations peuvent être communiquées simultanément.

1.4. LE GESTE DE LA LANGUE DES SIGNES

On a pu constater lors de l'étude précédente que les gestes de la LSF (Langue des Signes Française) et les gestes co-verbaux présentent une structuration et un fonctionnement similaires dans le cas de l'expression temporelle. Un geste co-verbal permet de transmettre plusieurs types d'information simultanément. C'est aussi le cas des gestes des langues des signes. Ces langues possèdent, comme toute langue, un vocabulaire et une syntaxe. Il est particulièrement intéressant d'étudier ce type de gestes car leur structure et leur utilisation est bien définie et ils se suffisent à eux-mêmes : leur interprétation ne dépend pas d'un message issu d'une autre modalité comme c'est le cas pour les gestes co-verbaux. La puissance d'expression de ces gestes incite à les prendre comme référence et à les étudier plus finement, afin de construire un outil de reconnaissance de gestes capable de prendre en compte toutes les potentialités du canal gestuel.

Après avoir présenté les données physiologiques qui différencient les langues gestuelles des langues orales, nous présentons la structuration des signes suivant les cinq paramètres couramment utilisés (configuration, mouvement, orientation, emplacement et mimique faciale). Quelques études de type phonologique[6] menées sur les langues des signes sont ensuite rapidement exposées.

1.4.1. LES DIFFERENCES ENTRE LES LANGUES ORALES ET LES LANGUES GESTUELLES

Trois données physiologiques différencient radicalement les langues des signes des langues orales :

- Une étude réalisée sur des signes isolés de l'ASL (American Sign Language) et des mots de l'américain parlé [Bellugi U. et Klima E. 1979] a montré que :
  - pour un même concept, le temps d'émission moyen d'un signe isolé est approximativement deux fois plus long que celui d'un mot,
  - à contenu et quantité d'information égale, le temps d'émission d'un discours en langue des signes et en américain parlé est approximativement le même.

6 La *phonologie* est la science qui étudie les sons du langage du point de vue de leur fonction dans le système de communication linguistique [Dubois J., Giacomo M. et al. 1973].

- Le système auditif humain est adapté à la discrimination temporelle tandis que le système visuel est adapté à la discrimination spatiale [Mc Neill D. 1992].
- Le signal analysé par l'oeil correspond directement au mouvement des articulateurs (les mains, les bras…), tandis que l'oreille n'analyse que l'effet sonore produit par le mouvement des articulateurs (les cordes vocales).

Selon Christian Cuxac [Cuxac C. 1983] ces différences conditionnent le fonctionnement des langues des signes. Les niveaux syntaxiques et sémantiques sont structurés de façon à "rattraper le temps perdu" en ce qui concerne l'émission de signes isolés de tout contexte. Cela se fait au moyen d'une spatialisation des rapports syntaxiques ainsi que d'une utilisation fréquente d'informations simultanées.

Ces informations sont portées par des **paramètres** de type phonologiques qui sont la configuration (la forme de la main), le mouvement de la main, l'emplacement et l'orientation de la main par rapport au corps et la mimique faciale [Moody B. 1983]. Parfois les deux mains interviennent et parfois une seule est utilisée.

Plusieurs informations hétérogènes peuvent être émises simultanément selon qu'elles font varier l'un ou l'autre de ces paramètres. On peut imaginer alors toute l'économie de temps réalisée si ces cinq paramètres sont émis simultanément et si chaque paramètre est exploité au niveau syntaxique.

Notons que d'autres parties du corps participent lors de l'émission d'un message en LSF, en particulier les épaules et le buste. Un geste en langue des signes représente un ensemble des mouvements effectués par différentes parties du corps simultanément. Mais seuls les paramètres cités précédemment ont été codifiés.

1.4.2. LES PARAMETRES

Les paramètres sont spatiaux et co-occurrents. Ils possèdent une valeur sémantique et une fonction syntaxique autonomes [Cuxac C. 1987]. L'exploitation syntaxique de ces paramètres tient compte de l'exigence d'iconicité des langues des signes. Ainsi, les différents paramètres sont spécialisés en fonction de leur nature même. Un aperçu de quelques fonctionnalités possibles pour les différents paramètres est donné ci-après. Une description plus complète est donnée au Chapitre 2.

1.4.2.1. La configuration

La **configuration** (forme de la main) reprend une partie de la forme de l'objet de l'action [Cuxac C. 1983].

Par exemple, selon la configuration utilisée, une action est appliquée à un instrument différent. La Figure 1.14 illustre le signe [**boire**] générique. En modifiant uniquement la forme de la main, on peut faire référence à des formes de récipient différentes. Par exemple, on peut exprimer le fait de boire dans un verre (configuration "**c**"[7] avec une seule main), un bol (configuration "**boule**" avec les deux mains), ou encore une tasse (configuration "**pince**" avec une seule main) (Figure 1.15).

Figure 1.14 : Le signe [**boire**][8].

Figure 1.15 : Les configurations "**c**", "**boule**" et "**pince**", applicables au verbe [**boire**].

7 Les configurations de la main sont référencées à l'aide de symboles ou de noms n'ayant pas de rapport avec leur signification. Ces labels sont dérivés de l'alphabet gestuel, de nombres ou de mots exprimant leur forme.

8 Les figures représentant des signes sont toutes extraites du dictionnaire de Bill Moody [Moody B. 1986] (© IVT).

1.4.2.2. Le mouvement

Le **mouvement** de la main représente une action exercée par ou sur l'objet. De plus, la dynamique du mouvement permet de différencier les aspects du verbe [Cuxac C. 1983].

Par exemple, le verbe [**voir**] (Figure 1.16) peut comporter les marques aspectuelles suivantes :

Mouvement	Paraphrase
bref	je vois brièvement
bref, petit et latéral	je vois en cachette
lent et grand	je vois longtemps
répétition	je vois souvent
répétition et bref	je vois "saccadé"
circulaire répété	je vois toujours
bref interrompu	j'ai presque vu

Tableau 1.3 : Exemples de marques aspectuelles pour le verbe voir.

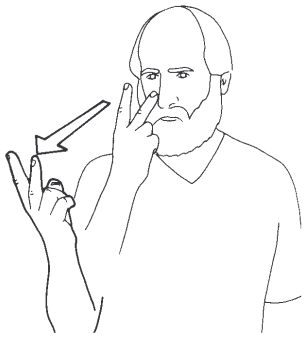


Figure 1.16 : Le signe [**voir**].

1.4.2.3. L'emplacement

L'**emplacement** de la main représente l'endroit où l'action est effectuée ou un endroit générique [Cuxac C. 1983].

Par exemple, un verbe comme [**opérer**] possède un emplacement générique, mais peut être réalisé à d'autres endroits du corps, pour indiquer une opération spécifique : opérer du ventre, de l'oeil, du coeur… (Figure 1.17)

Figure 1.17 : Le signe [**opérer**], effectué sur le buste.

1.4.2.4. L'orientation

L'**orientation** permet de conjuguer certains verbes, ou de préciser l'orientation d'objets [Cuxac C. 1983].

En langue des signes, il existe deux classes de verbes : les verbes **directionnels** et les verbes **non directionnels**. Les verbes directionnels se conjuguent dans l'espace. Les différents intervenants du discours sont positionnés dans un repère fictif ayant le signeur comme origine. Puis la direction du mouvement et l'*orientation* de la main déterminent l'agent et le patient.

C'est le cas de verbes tels que donner, envoyer, demander… Dans la Figure 1.18, l'agent est le signeur et le patient est l'interlocuteur, car le mouvement est dirigé du signeur vers son interlocuteur. La traduction complète du signe est [**je t'envoie** (quelque chose)]. En modifiant la direction du mouvement ainsi que les orientations de début et de fin de geste, on modifie la conjugaison du verbe.

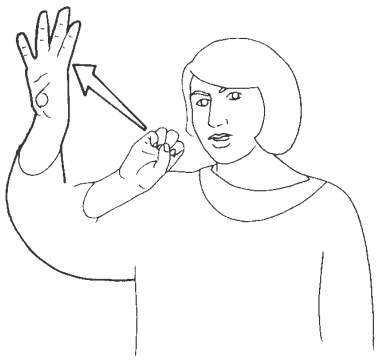


Figure 1.18 : Le signe [**je t'envoie**].

Dans le cas des verbes qui ne sont pas directionnels, l'orientation n'a pas de fonction syntaxique particulière (voir par exemple le verbe [**opérer**], Figure 1.17).

1.4.2.5. La mimique faciale

Enfin, la **mimique faciale** permet d'exprimer le mode du discours (interrogatif, négatif…) ou joue le rôle des compléments de manière (plaisir, colère, envie, gêne…). Ces compléments de manière permettent aussi d'exprimer le point de vue du signeur sur le contenu de l'énoncé.

Certains signes ne se différencient que par la mimique faciale, comme par exemple les signes [**impossible**] et [**détester**] (Figure 1.19).

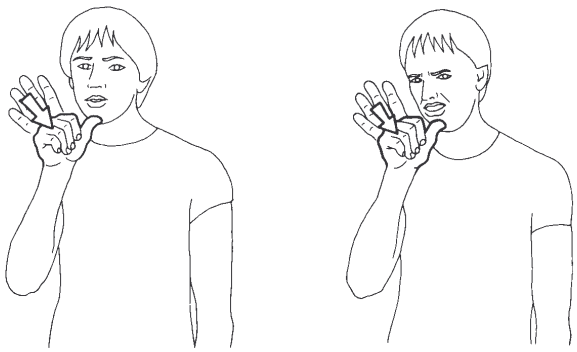


Figure 1.19 : Les signes [**impossible**] à gauche et [**détester**] à droite.

1.4.2.6. Un exemple

L'ensemble de ces paramètres se combinent pour former un geste. Par exemple, le geste représenté dans la Figure 1.20 signifie [**regarde-moi !**]. Il est constitué des paramètres suivants :

- La configuration indique que l'objet de l'action est composé de deux éléments situés côte à côte.
- Le mouvement de la main (a) représente une action simple (une droite) dirigée vers le signeur.
- L'emplacement indique que l'action est réalisée au niveau des yeux.
- L'orientation (et la direction du mouvement) indique que la personne placée devant le signeur est l'agent de l'action.
- La mimique indique le mode impératif par le biais d'une expression insistante dans les yeux et d'un mouvement ferme de la tête (b).

De plus, le mode impératif est confirmé par une accentuation tonique sur le verbe : le signe est réalisé rapidement et avec un mouvement plus ferme qu'usuellement.

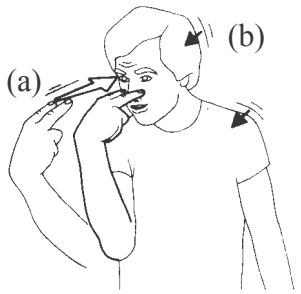


Figure 1.20 : Le signe [**regarde moi !**].

On peut voir la puissance d'un tel mode de communication, où en un seul geste, il est possible de trouver plusieurs couches d'informations : en plus de la coexistence des cinq paramètres, l'analyse dynamique des gestes donne des informations syntaxiques et sémantiques supplémentaires.

1.4.3. PHONOLOGIE

Ce n'est que depuis les années soixante que les langues des signes sont considérées comme des langues à part entière au même titre que les langues orales. Un linguiste américain, W.C. Stokoe, a démontré pour la première fois que L'ASL, American Sign Language, possédait le principe de double-articulation[9] [Stokoe W. 1960] Dans le cadre d'études "phonologiques"[10], il a proposé des unités minimales différentielles pour l'ASL. Ces unités minimales, qu'il appelle *chérèmes*, correspondent aux phonèmes dans les langues orales. Elles se combinent entre elles pour former des *kinèmes*, correspondant aux monèmes dans les langues orales.

Les études sur les langues des signes sont encore peu nombreuses. La plupart sont des études américaines, dans la mouvance de celle de Stokoe. En France, pour diverses raisons historiques (voir Annexe 1.2), la LSF a été interdite pendant un siècle. Encore maintenant, certains linguistes ne la reconnaissent pas comme une véritable langue. C'est sans doute pourquoi peu de chercheurs français s'y intéressent. Cependant, toujours pour des raisons historiques, l'ASL et la LSF sont très proches et les études faites sur l'une des langues peuvent servir de base pour étudier l'autre, tout du moins en ce qui concerne le vocabulaire.

Les études réalisées à ce jour sont surtout de type "phonologique" [Lane H., Boyes-Braem P. et al. 1976], [Sandler W. 1989], [Perlmutter D. M. 1990]. Un des buts recherché est de créer un alphabet qui permettrait de transcrire un discours en langue des signes à l'aide d'un système graphique.

Trois problèmes se posent dans le cas des langues des signes. D'une part, plusieurs parties du corps interviennent pour exprimer des informations simultanées, d'autre part, les

---

9 La double-articulation est l'organisation spécifique du langage humain selon lequel tout énoncé s'articule sur deux plans [Dubois J., Giacomo M. et al. 1973]. Le premier plan correspond à l'articulation de l'énoncé en unités douées de sens dont les plus petites sont appelées monèmes ou morphème. Le deuxième plan correspond à l'articulation de chaque monème en unités dépourvues de sens dont les plus petites sont les phonèmes.

10 La *phonologie* est la science qui étudie les sons du langage du point de vue de leur fonction dans le système de communication linguistique. La *phonétique* étudie les sons du langage dans leur réalisation concrète, indépendamment de leur fonction linguistique. Exemple : le mot "mer" possède des traits phonologiques fixes ("liquide" et "non-latéral"), et des traits phonétiques variables selon la prononciation (dentale roulée / accent bourguignon ; vélaire roulée / accent grasseyé ; vélaire constrictive / accent parisien) [Dubois J., Giacomo M. et al. 1973].

informations sont données par des valeurs qui peuvent être dynamiques, comme par exemple le mouvement du bras, ou le mouvement des doigts. Enfin, ces informations sont spatiales.

Pour simplifier le problème, certains chercheurs proposent des systèmes de représentation dans lesquels les signes sont segmentés en unités statiques parallèles, qui se succèdent dans le temps.

Citons à titre d'exemple, le modèle *Movement Hold*, qui est un modèle de segmentation phonologique [Liddel S. K. 1990]. Il propose une structuration des signes en deux couches :

- une couche dite "segmentale", dans laquelle on segmente le signe en une succession de segments de mouvement (M) et de tenue (T). Le segment de mouvement correspond à des variations articulatoires et le segment de tenue correspond à un état constant des articulateurs. C'est dans cette couche que sont spécifiés le type de mouvement et la manière dont il est exécuté ;
- une couche articulatoire, dans laquelle sont décrites des successions de valeurs pour chaque articulateur, comme la configuration de la main, l'emplacement, l'orientation et les points de contact, dans des sous-couches distinctes et indépendantes, ce qui permet d'éviter les informations redondantes.

En ce qui concerne plus précisément le paramètre de mouvement, selon les études réalisées, les mouvements des doigts, du poignet et du bras sont distingués ou pas. Par exemple, dans une étude portant sur la LSF, Monique Touati considère le paramètre *signation*, qui correspond à l'ensemble des mouvements (bras, main et poignet) effectués durant le signe [Touati M. 1983]. Elle définit une liste de *SIG* (unité minimale de SIGnation), qui peuvent s'appliquer à tous les paramètres. Par exemple, le SIG **VB**, qui signifie "vers le bas", peut s'appliquer à un mouvement du bras, à une orientation de la paume de la main, ou encore à une localisation (par exemple, l'emplacement "menton" est codé par "partie du visage, vers le bas").

Ces études ont pour finalité de décomposer au maximum un signe de manière à pouvoir le codifier, le représenter et l'analyser. Dans le cadre d'un outil de reconnaissance de gestes, une telle démarche ne convient pas car elle ne tient pas compte de la réalisation effective du signe et des articulateurs mis en jeu.

C'est pourquoi nous avons dû mener notre propre étude, qui est présentée dans le Chapitre 2.

1.5. CONCLUSION

L'étude des gestes co-verbaux et des gestes de la LSF nous a conduit à constater deux faits importants :

- Le geste de la main est structuré en quatre paramètres qui sont la configuration, le mouvement, l'orientation et l'emplacement de la main. Chacun d'eux est chargé de véhiculer une information d'un type différent. Ces quatre types d'informations peuvent être émis simultanément.

  Remarquons que pour l'application test réalisée dans le cadre des interfaces gestuelles, ce type de spécialisation des parties différentes de la main pour différents types d'informations a été utilisée : le mouvement sert à véhiculer la commande, c'est-à-dire le verbe, tandis que la configuration sert à véhiculer les paramètres, c'est-à-dire le sujet et le complément du verbe.
- Le locuteur ou le signeur utilise une scène de narration placée devant lui et c'est en plaçant des objets, des personnes ou des événements au sein de cette scène qu'il construit l'image contenant l'information qu'il veut communiquer. C'est le moyen utilisé pour représenter le contexte, afin de résoudre l'ambiguïté du message.

La similitude entre les gestes co-verbaux et les gestes de la LSF nous pousse à mener une étude plus approfondie de ces derniers afin de proposer un outil de reconnaissance et de compréhension tenant compte de l'ensemble des propriétés de simultanéité d'information et de structuration de l'espace scénique exploitées en LSF.

Cette étude doit permettre d'intégrer un outil de reconnaissance et de compréhension dans des applications multimodales où la parole n'est pas prédominante ou pas disponible.

Chapitre 2

ÉTUDE DES PARAMETRES DE LA LSF

Comme nous l'avons vu au chapitre précédent, un geste permet de transmettre plusieurs informations simultanément par l'intermédiaire de paramètres qui transportent en parallèle des informations de types différents. Avant de proposer un système informatique adapté au canal gestuel, il faut étudier plus finement comment sont construits et utilisés ces paramètres. Nous avons choisi d'étudier les gestes de la LSF, car leur structure syntaxique et sémantique est bien définie et leur interprétation ne dépend pas d'une autre modalité.

Pour ce faire nous avons réalisé et exploité une base de données contenant la description des signes contenus dans le premier tome du dictionnaire de Bill Moody. Cette étude a pour rôle d'informer sur les principes de construction et d'utilisation des signes afin de déterminer l'architecture d'un système de reconnaissance et de compréhension dédié au canal gestuel. Elle permet aussi d'aider à la construction de corpus de gestes pour évaluer ce système.

Après avoir indiqué comment l'étude a été réalisée, notre définition des quatre premiers paramètres qui constituent un signe en LSF est précisée. Puis pour chaque paramètre, les champs de la base de données qui ont été définis et les principaux résultats obtenus grâce à l'exploitation de cette dernière sont présentés. En fin de chapitre, la structure de la base de données contenant la description des signes est rappelée, avec un exemple complet de description d'un signe. Le dernier paragraphe contient une synthèse des principaux résultats.

2.1. INTRODUCTION

Le travail réalisé a consisté à observer et à décrire une par une les images représentant les différents signes contenus dans le premier tome du dictionnaire de Bill Moody [Moody B. 1986], soit 1359 signes. Parmi ces signes, certains (102) ont été considérés comme étant composés de deux ou plusieurs signes, comme par exemple le signe [**après**] [**midi**], montré Figure 2.1. Ces signes n'ont pas été pris en compte, car ils sont équivalents à une séquence de signes simples et c'est sur un total de 1257 signes simples que porte l'étude.

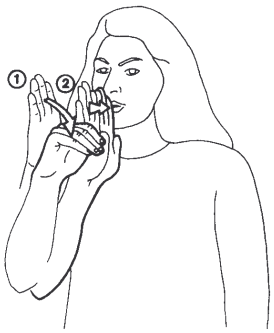


Figure 2.1 : Le signe [**après**] [**midi**].

Afin de pouvoir obtenir des informations chiffrées, toutes les descriptions ont été stockées dans un base de donnée (Paradox sur PC). Dans un premier temps, pour chaque paramètre, il a fallu déterminer les différents champs nécessaires pour stoker les informations. Il a fallu ensuite déterminer des systèmes de notations, en particulier pour représenter les types de mouvement, leur orientation, etc. L'étape suivante a consisté à saisir toutes les descriptions dans les différents champs de la base de donnée. Enfin, l'exploitation de la base de donnée, par l'intermédiaire d'une série de requêtes, a permis d'obtenir toute une quantité d'informations chiffrées. Les valeurs numériques précises sont présentées en Annexe 2 (de 2.1 à 2.5).

La description des signes s'est faite paramètre par paramètre, en prenant appui sur les définitions données au paragraphe suivant.

2.2. DEFINITION DES PARAMETRES

Notre propos n'est pas de trouver un système de représentation de la langue des signes française, mais plutôt de déterminer quelles sont les unités minimales des signes qui peuvent être prises en compte dans un système de reconnaissance de gestes. C'est d'une étude de type "articulatoire" plutôt que d'une étude de type phonologique qui est nécessaire. Une telle étude n'ayant pas encore été réalisée à ce jour, nous avons dû la mettre en oeuvre.

Nous nous sommes intéressée aux mouvements de parties du corps (mains et bras) lors de l'émission du message gestuel, indépendamment de leur rôle linguistique. De ce point de vue, les muscles qui permettent de faire bouger les doigts sont différents de ceux qui font se déplacer le bras [Calais-Germain B. 1989]. De plus, pour un même déplacement du bras, il est possible, à l'aide de muscles spécifiques, d'effectuer une rotation ou une flexion du poignet. Ainsi, contrairement à ce qui est souvent considéré dans les études phonologiques (Chapitre 1, Paragraphe 1.4.3), pour nous le mouvement est décomposé en mouvement des doigts, de la main, du bras. D'où les définitions suivantes, adaptées à l'étude des gestes de la main :

- La **configuration** représente la forme et le mouvement des doigts de la main. Les doigts peuvent être immobiles ou pas, donc la configuration peut être statique ou dynamique.
- L'**orientation** est la valeur de deux directions, celle de l'axe de la main et celle de la paume, ainsi que l'état de l'articulation poignet (repos, rotation ou flexion). Ces valeurs peuvent varier durant l'exécution du signe. Ce paramètre peut être statique ou dynamique.
- Le **mouvement** représente la trajectoire du déplacement de l'extrémité de l'avant-bras (côté poignet). Lorsque le bras est immobile, on parlera de paramètre mouvement "statique".
- L'**emplacement** représente la zone dans laquelle le signe est effectué par rapport au signeur. A chaque fois que le bras bouge, l'emplacement varie. Il y a une corrélation directe entre le paramètre de mouvement et le paramètre d'emplacement. Ce paramètre peut être statique ou dynamique.

Nous cherchons à énumérer, pour chaque paramètre, tous les types de comportement pouvant exister, à partir de la perception visuelle de ces quatre paramètres dans l'espace. Par exemple, nous décrivons le signe [**ver**], désignant un ver de terre (Figure 2.2), vu du signeur, de la manière suivante (seule la main droite est décrite) :

- La configuration est dynamique, il s'agit de la fermeture de l'index qui passe d'une position tendue à une position mi-pliée. Cette configuration est répétée plusieurs fois ;
- L'orientation est composée des informations suivantes : l'axe de la main est dirigé vers la gauche, la paume est dirigée vers le bas et le poignet est tendu ;
- Le mouvement est dynamique, il s'agit d'un déplacement simple dont la trajectoire est une droite horizontale allant de droite à gauche ;
- L'emplacement est une zone neutre située face au signeur.

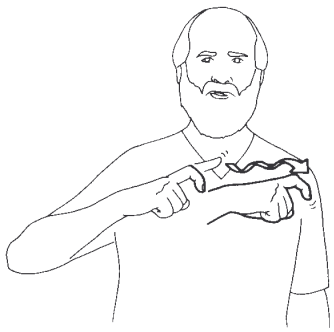


Figure 2.2 : Le signe [**ver**].

Ainsi, la spécificité de cette description est que le mouvement de l'index est dissocié du mouvement du bras, à la différence de la description donnée dans le dictionnaire [Moody B. 1986], dans lequel le mouvement est représenté par un seul symbole : une vaguelette horizontale, concaténation des mouvements de l'index et du bras.

2.3. LES QUATRE PARAMETRES

Ce paragraphe comporte, pour chaque paramètre, deux parties. La première explique le choix des différents champs de la base de données pour ce paramètre. La deuxième contient une analyse des résultats des requêtes concernant ce paramètre. Cette analyse est parfois complétée par une description de la fonction syntaxique et du rôle sémantique du paramètre.

2.3.1. CONFIGURATION

2.3.1.1. Base de données

Notations

Dans la suite du mémoire, les configurations seront la plupart du temps illustrées par un dessin. Cependant à chaque configuration est assigné un nom arbitraire, qui est le plus souvent issu de la dactylologie (exemples : **c**, **o**), des chiffres (exemples : **0**, **1**), ou d'une évocation brève de la forme de la main (exemples : **index**, **bec5**). Ces notations ne sont aucunement liées au sens des signes. elles sont inspirées de celles utilisées dans le dictionnaire de Bill Moody et ont été complétées afin de couvrir tous les cas rencontrés. La liste des configurations statiques est présentée en Annexe 2.1.

Pour chaque geste, la base de données contiendra un champ identification nommé **Config1.**

Configurations dynamiques

Comme nous allons le voir par la suite, une configuration peut être dynamique. Une telle configuration est composée d'une séquence de configurations statiques distinctes. Dans ce cas, elle est notée de manière simplifiée par la configuration initiale suivie de la configuration finale. En effet, toutes les configurations intermédiaires sont déductibles de ces deux configurations extrêmes, la variation se faisant d'une manière continue.

Par exemple, pour les signes [**attraper**] et [**éponge**] (Figure 2.3), la configuration est notée **5/s**, où **5** représente la configuration initiale (main ouverte) et **s** représente la configuration finale (main fermée).

Comme nous allons le voir dans la partie *Analyse*, plusieurs types de comportement dynamique existent pour la configuration (ouverture, fermeture, vibration…). Le code utilisé

pour noter les configurations ne suffit pas pour distinguer tous les cas. Par exemple, une configuration notée **boule** peut être associée à un comportement statique ou à une vibration des doigts, comme pour le signe [**araignée**] Figure 2.4.

Pour chaque geste, la base de données contiendra un champ spécifiant son type de comportement, nommé **Type Config**. Cela permet de simplifier ensuite les requêtes relatives à la configuration.

Répétition

Dans le cas d'une configuration dynamique, il peut arriver que le mouvement des doigts soit répété. Il faut spécifier la présence éventuelle d'une répétition pour la configuration.

Par exemple, pour le signe [**éponge**] (Figure 2.3), la configuration est notée **5/s**, comme pour le signe [**attraper**]. Cependant, on les distingue en indiquant que la configuration est répétée pour le signe [**éponge**].

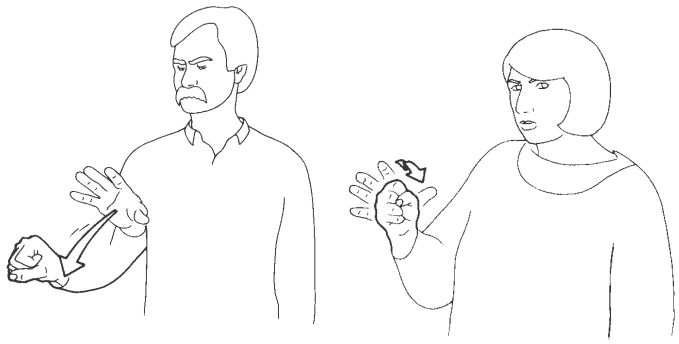


Figure 2.3 : Les signes [**attraper**] et [**éponge**].

Pour chaque geste, la base de données contiendra un champ spécifiant la répétition éventuelle de la configuration si elle est dynamique, nommé **Répèt Config**.

Main dominante, main dominée

Les signes peuvent être réalisés à l'aide d'une seule main ou des deux mains. Lorsque les signes sont réalisés à l'aide des deux mains, on distingue ces dernières d'après leur rôle fonctionnel, plutôt qu'en différenciant le coté gauche du coté droit. Ainsi, on parle de la **main dominante** et la **main dominée**. Comme ces termes le laissent entendre, la main dominante a un rôle plus actif que la **main dominée**, dans l'exécution du signe d'une part, dans la fonction grammaticale d'autre part. La Figure 2.4 montre le signe [**araignée**], dans lequel la main

dominante à un rôle actif - elle montre la forme et les mouvements de l'araignée - tandis que la main dominée sert de repère locatif fixe par rapport à la main dominante et est immobile. En général, la main dominante est la droite pour les droitiers et la gauche pour les gauchers.

Figure 2.4 : Le signe [**araignée**].

Comme le montre la Figure 2.4, main dominante et main dominée peuvent posséder des configurations distinctes qui devront toutes les deux être spécifiées. Dans la base de données, la configuration de la main dominante est stockée dans le champ **Config1** et celle de la main dominée dans le champ **Config2**.

Afin de simplifier les requêtes relatives aux deux configurations, un champ permettant de spécifier les différences entre les deux configurations, nommé **C1C2**, a été ajouté.

Liste des champs

En résumé, pour le paramètre de configuration, la base de données contient les cinq champs suivants :

- **Config 1** : Configuration de la main dominante.
- **Config 2** : Configuration de la main dominée.
- **Type Config** : Type de comportement dynamique de la main dominante.
- **Répèt Config** : Répétition éventuelle en cas de configuration dynamique de la main dominante.
- **C1C2** : Différences entre les configurations des 2 mains, le cas échéant.

Les différentes valeurs possibles pour chacun de ces champs est donnée à la fin du chapitre. Le paragraphe Analyse indique les valeurs les plus fréquentes.

Exemple

Exemple de codage des configurations sur le signe [**araignée**] (Figure 2.4) :

- **Config 1** : boule
- **Config 2** : pince
- **Type Config** : vibration
- **Répèt Config** : non
- **C1C2** : différent

2.3.1.2. Analyse

Ce paragraphe présente les résultats de l'analyse effectuée sur le paramètre de configuration. Trois résultats sont exposés. Le premier est une liste des différentes configurations utilisées dans le corpus. Ce résultat est complété par une étude des gestes non standard et plus particulièrement les classificateurs. Le deuxième résultat porte sur les configurations dont le comportement est dynamique. Enfin, les rapports entre les configurations des deux mains sont présentés.

Liste des configurations

L'étude réalisée sur le dictionnaire de Bill Moody a permis de constater qu'il y a une très grande diversité de configurations de la main (139). Nous les avons répertoriées et avons calculé les pourcentages d'occurrence pour chacun d'entre eux dans le dictionnaire (voir Annexe 2.1). Ceci doit permettre de connaître le taux de couverture de notre système de reconnaissance, par rapport au corpus étudié, en fonction des configurations qu'il sera en mesure de reconnaître. Il ne s'agit pas d'un taux de couverture réel puisqu'on ne connaît pas la fréquence d'utilisation des signes dans une conversation en LSF, mais cela permet d'avoir une première estimation.

Cette estimation porte sur des gestes qui sont répertoriés dans un dictionnaire. Ces gestes sont appelés signes **standard** (ou conventionnels). Chaque langue des signes possède un vocabulaire de signes standard spécifique à chaque pays.

Mais un second type de signes est très fréquemment utilisé dans les dialogues en langue des signes : l'ensemble des signes **non standard**. Ils sont créés durant le discours, en fonction des besoins et du contexte. De ce fait, ils ne peuvent pas être répertoriés dans un dictionnaire. Ils sont basés sur les caractéristiques physiques des objets et ainsi, ils peuvent être interprétés

indépendamment de la culture du pays. Ces signes sont utilisés durant les conversations entre sourds de pays différents. Mais ils sont aussi souvent utilisés dans les conversations ordinaires entre sourds d'un même pays. Un exemple de signe non standard est le classificateur.

Les classificateurs

Un **classificateur** est un signe qui décrit et représente toute une classe (ou famille) d'objets ayant une forme, une taille ou une épaisseur similaire [Moody B. 1983]. Ils peuvent avoir deux fonctions syntaxiques :

- *Fonction descriptive*
  - Description de la forme des objets.

    Trois exemples de classificateurs pour décrire des objets longs sont illustrés Figure 2.5.

Figure 2.5 : (1) très fin, (2) plutôt fin, (3) épais, pour un objet long et horizontal.

  - Localisation d'objets dans l'espace.

    Dans l'exemple présenté Figure 2.6, le signe [**verre**] est suivi du classificateur **petit-récipient**, pour spécifier la localisation du verre dans la scène.

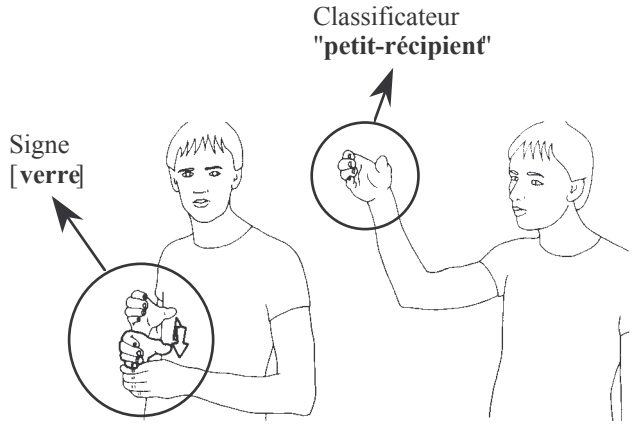


Figure 2.6 : Phrase "**le verre là-haut**".

- *Fonction de remplacement*.

  Dans ce cas, ils jouent le rôle de "super-pronom", car ils remplacent le référent (comme les pronoms), mais en plus ils montrent sa forme.

  - Il est possible d'incorporer dans le verbe un super-pronom qui a la fonction de complément d'objet. C'est le cas de la main dominée dans l'exemple présenté Figure 2.8 (une main plate qui représente par exemple une table).
  - Il est possible d'incorporer dans le verbe un super-pronom qui a la fonction de sujet. C'est le cas de la main dominante dans l'exemple présenté Figure 2.8 (une main en forme de C pour représenter par exemple un verre qui tombe).

    Ainsi, pour exprimer qu'un objet de type petit récipient (par exemple un verre) est tombé d'un objet de type surface plane (par exemple une table), plutôt que de réaliser la séquence de trois gestes montrée Figure 2.7, le signeur exécute un unique geste, montré Figure 2.8.

Figure 2.7 : Classificateur **petit-récipient**, signe [**tomber**], classificateur **surface.**

Figure 2.8 : Phrase "**il en tombe**" (par exemple : il = le verre, en = de la table).

- Il est possible d'incorporer le classificateur dans un verbe de déplacement (exemple Figure 2.9)

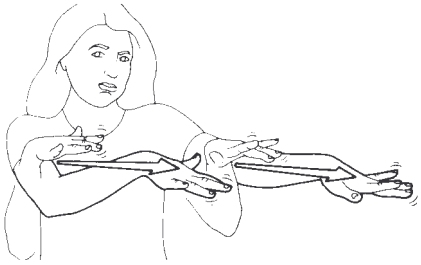


Figure 2.9 : Phrase "**la foule avance**".

- Il est possible d'utiliser un classificateur pour décrire la localisation de personnes. Par exemple, comme le montre la Figure 2.10, la phrase "deux personnes s'assoient face à face aux bouts d'une longue table" est composée d'une séquence de trois gestes : le signe [**table**], un classificateur **objet-plat-long-et-horizontal**, un classificateur **deux-personnes-face-à-face-qui-s'assoient**.

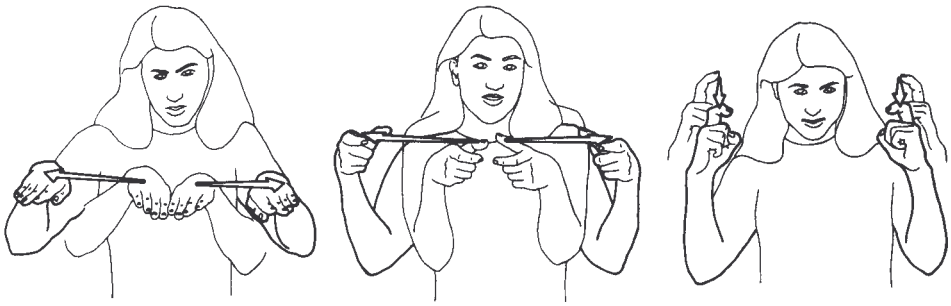


Figure 2.10 : Phrase "Deux personnes s'assoient face à face aux bouts d'une longue table".

Parmi tous les classificateurs existants, les plus utilisés sont les suivants :

- Des classificateurs qui servent à décrire des objets ronds. La configuration peut être **c**, **boule** ou **0.**

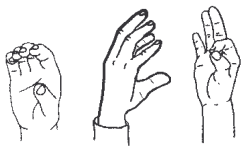


Figure 2.11 : Classificateurs pour objets ronds.

- Des classificateurs qui servent à décrire des objets plats. Typiquement, les configurations **plat**, **moufle** ou **2**.

Figure 2.12 : Classificateurs pour objets plats.

- Des classificateurs qui servent à décrire des personnes. Par exemple à l'aide des configurations **index**, **s**, **5**,**v** ou **1**.

Figure 2.13 : Classificateurs pour personnes.

L'utilisation de tels signes est très courante et ils ne sont pas répertoriés dans le dictionnaire, puisque leur sens dépend entièrement du contexte. Les configurations utilisées pour les classificateurs sont parmi celles qui sont le plus rencontrées dans le dictionnaire. Ce sont celles dont la forme est le mieux à même de rappeler une caractéristique morphologique simple d'un objet ou d'une personne.

Ce sont en priorité ces configurations qu'il est important de reconnaître dans une application de type reconnaissance de gestes de la LSF car d'une part, elles couvrent près de la moitié des signes du corpus (voir Annexe 2.1) et d'autre part, en tant que classificateurs, il est sûr que les fréquences d'utilisation de telles configurations dans des conversations réelles sont élevées.

Comportement dynamique

Ce paragraphe présente une classification des configurations dynamiques, basée sur les différents comportements dynamiques observés dans le corpus.

Les configurations sont différenciées selon des critères de variation des vitesses angulaires observées pour chaque doigt (provoquant une augmentation ou diminution des valeurs angulaires). La valeur angulaire représente l'état de flexion d'une articulation (typiquement entre 0° et 90°). Les différents types de comportement dynamique observés sont les suivants :

- *Statique* : les valeurs angulaires de toutes les articulations ne varient pas durant le signe ;
- *Fermeture :* la valeur angulaire d'au moins une articulation augmente durant le signe ;
- *Ouverture :* la valeur angulaire d'au moins une articulation diminue durant le signe ;
- *Vibration :* au moins un doigt possède un mouvement répété de flexion/extension de faible amplitude (quelques degrés) durant le signe (ex : signe [**araignée**], Figure 2.4) ;
- *Frottement :* certaines parties de la main frottent contre d'autres.

D'autres comportements, correspondant à une composition de plusieurs configurations, statiques ou dynamiques, ont été observés :

- *Fermeture/Ouverture :* une configuration dynamique de type fermeture est suivie d'une configuration dynamique de type ouverture ;
- *Ouverture/Fermeture :* une configuration dynamique de type ouverture est suivie d'une configuration dynamique de type fermeture ;
- *Ouverture/Vibration :* une configuration dynamique de type ouverture est suivie d'une configuration dynamique de type vibration.

Il n'existe pas de cas où coexistent dans un même signe des variations opposées : il n'existe pas de cas où certains doigts se plient et d'autres se déplient simultanément (dans le

tome 1 de Moody [Moody B. 1986], seul existe le signe [**ou**], qui est un signe méthodique[1], composé d'un **o** suivi d'un **u**).

Figure 2.15 : Le signe [**ou**].

La répartition des signes du corpus parmi ces types de comportement a été étudiée (Figure 2.16). Près de 80 % des signes possèdent une configuration de type *statique* (les doigts ne bougent pas). Par ailleurs, 12 % des signes possèdent une configuration dynamique de type *fermeture* (un ou plusieurs doigts se plient durant le signe) et 6% possèdent une configuration dynamique de type *ouverture* (un ou plusieurs doigts se déplient durant le signe).

Moins de 3 % des signes font partie d'une des cinq classes restantes (34 signes). Ces cinq classes correspondent à des mouvements des doigts plus complexes d'un point de vue articulatoire.

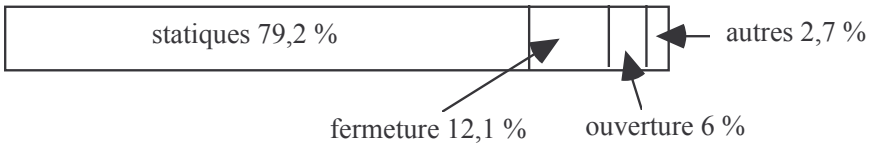


Figure 2.16 : Les classes de configurations les plus fréquentes.

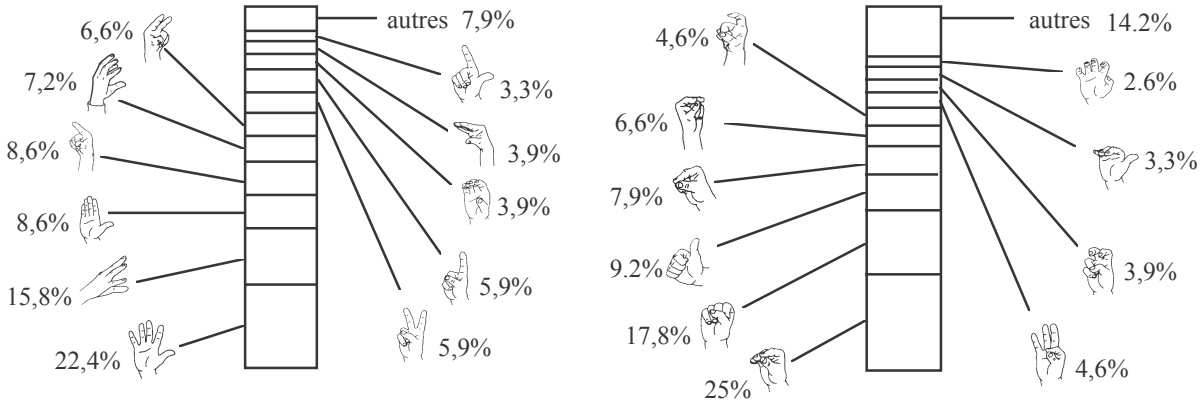
Il semble au premier abord que les configurations les plus compliquées, du point de vue réalisation comme du point de vue interprétation visuelle, sont utilisés le moins possible. Par conséquent, une analyse plus détaillée des deux classes fermeture et ouverture a été menée.

---

1 Les signes dits méthodiques sont des signes artificiels inventés par l'Abbé de l'Epée pour enseigner le français aux enfants sourds. Ils sont restés dans le lexique, mais sont rarement utilisés par les sourds, car ils sont étrangers à la grammaire gestuelle [Moody B. 1983].

Classe fermeture

Cette classe couvre environ 12 % de toutes les configurations rencontrées dans le dictionnaire de Bill Moody, soit 152 signes. Les configurations de début et de fin de geste les plus souvent rencontrées parmi ces 152 signes ont été étudiées. Les résultats sont illustrés dans la Figure 2.17.



a. configuration de début

b. configuration de fin

Figure 2.17 : Les configurations de début et de fin les plus fréquentes pour la classe fermeture.

Remarques :

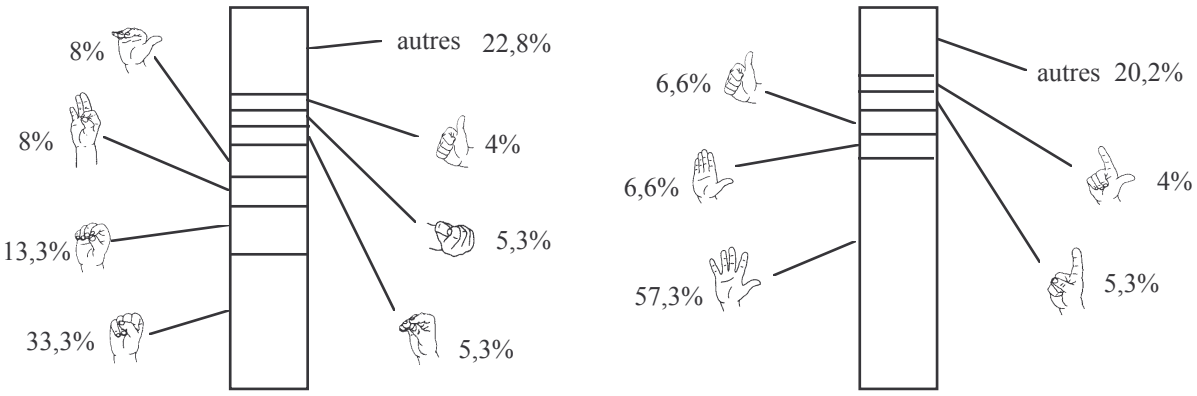
- Pour les trois configurations de début de geste les plus fréquentes, les cinq doigts sont tendus. Ce sont les configurations les plus ouvertes.
- Presque toutes les configurations de fin de fermeture sont celles pour lesquelles le mouvement des doigts est stoppé à cause d'un contact avec un doigt ou une partie de la main. Par exemple, le pouce permet de stopper le mouvement des autres doigts dans le geste finissant par la configuration **bec5** (Figure 2.18).

Figure 2.18 : Configuration **bec5.**

Classe ouverture

Cette classe couvre 6 % de toutes les configurations rencontrées dans le dictionnaire de Moody, soit 75 signes. Les configurations de début et de fin de geste les plus souvent

rencontrées parmi ces 75 signes ont été répertoriées. Les résultats sont illustrés dans la Figure 2.19.



a. configuration de début

b. configuration de fin

Figure 2.19 : Les configurations de début et de fin les plus fréquentes pour la classe ouverture.

Remarques :

- Pour cette classe, deux configurations représentent à elles seules plus de 40 % des configurations de début. Il s'agit du **s** (Figure 2.21) et du **o**, qui sont les configurations les plus fermées.
- Quant aux configurations de fin, la configuration **5** (Figure 2.20) représente à elle seule près de 60 % des cas rencontrés.

Liens entre les classes fermeture et ouverture

Quand ces deux classes, qui sont duales, sont étudiées simultanément, de nouvelles remarques peuvent être ajoutées :

- La configuration de début la plus fréquente pour la classe fermeture est identique à la configuration de fin la plus fréquente pour la classe ouverture (**5**). Il s'agit de la configuration la plus "ouverte", du point de vue articulatoire et la plus facile à réaliser, car tous les doigts sont identiquement tendus (Figure 2.20).

Figure 2.20 : Configuration **5.**

- La configuration de début la plus fréquente pour la classe ouverture (**s**) est la même que la deuxième configuration de fin la plus fréquente pour la classe fermeture. Il s'agit de la configuration la plus "fermée", du point de vue articulatoire et la plus facile à réaliser, car tous les doigts sont identiquement pliés (Figure 2.21).

Figure 2.21 : Configuration **s.**

- Par contre, la configuration de fin la plus fréquente pour la classe fermeture est **bec5**. Cette configuration est plus rapide à atteindre à partir d'une configuration ouverte que la configuration **s**. La distance que les doigts doivent parcourir est plus petite.

La conclusion tirée de l'ensemble de ces remarques est que la construction des configurations dynamiques semble être motivée par la facilité de production gestuelle pour le signeur.

Les résultats concernant ces deux classes de configurations ont été regroupés dans un graphe qui montre les variations les plus fréquentes (Figure 2.22). Cela a permis de faire de nouvelles constatations.

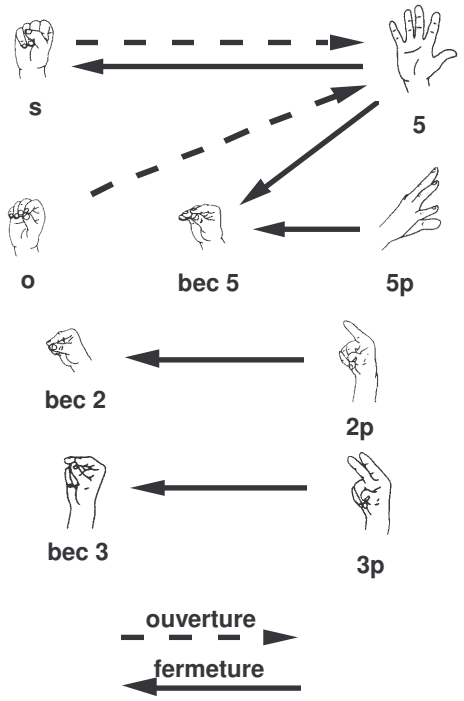


Figure 2.22 : Graphe des variations en fermeture et ouverture les plus fréquentes.

- La première remarque est que la variation en ouverture la plus fréquente (**s -> 5**) utilise les configurations **s** et **5** qui sont respectivement la plus fermée et la plus ouverte.
- Une autre remarque est que les variations d'ouverture et de fermeture les plus fréquentes sont celles pour lesquelles tous les doigts ont la même forme (par exemple, la fermeture **5 -> s** pour laquelle tous les doigts sont tendus puis pliés, ou l'ouverture **o -> 5** pour laquelle tous les doigts sont arrondis puis tendus).
- Certaines configurations sont utilisées uniquement en début de variation, ou uniquement en fin de variation. Elles correspondent toujours à des variation difficiles à réaliser précisément et rapidement (exemples : **o** est toujours utilisée en début d'ouverture, **bec2** est toujours utilisée en fin de fermeture).
- Certaines configurations, initiales ou finales, présentes dans les configurations dynamiques, sont rarement (parfois même jamais) rencontrées parmi l'ensemble des configurations statiques (exemples : **3p**, **2p**)
- Enfin, les variations, en ouverture ou en fermeture, concernent les ensembles de doigts suivants :
  - tous les doigts,
  - tous les doigts sauf le pouce,

- le pouce et l'index,
- l'index et le majeur,
- le pouce, l'index et le majeur.

Les annulaire et auriculaire ont un rôle passif. Soit ils sont statiques, soit ils suivent la même variation que les autres doigts. Ils n'ont jamais de comportement dynamique propre.

Nous avons pu remarquer que les configurations dynamiques les plus fréquentes sont celles pour lesquelles les variations sont importantes, sans doute pour permettre une discrimination des formes plus aisée. Ce sont aussi les plus simples. Globalement, plus l'aspect dynamique est complexe, plus la forme de la main est simple. Il semble que, du fait que la communication soit basée sur une analyse visuelle de plusieurs paramètres en parallèle, chacun des paramètres reste à un niveau limité de complexité visuelle.

Rapport entre les configurations des deux mains

Les observations effectuées sur les relations entre les deux mains sont les suivantes :

- Parfois les signes se font avec les deux mains (62,5% de cas), parfois avec une seule (37,5% des cas).
- Pour plus des trois quarts des signes qui utilisent les deux mains, les configurations sont identiques. Lorsque les configurations sont différentes, la configuration de la main dominée est simple et correspond en général à une des configurations utilisées dans les classificateurs.
- Les deux configurations sont différentes lorsque seul le bras de la main dominante est mobile. Lorsque les deux bras sont mobiles, les deux configurations sont identiques.

Ces informations pourront être utiles si l'on souhaite par la suite développer un système de reconnaissance du mouvement des deux mains.

Synthèse

L'étude réalisée sur le paramètre de configuration a permis de répertorier l'ensemble des configurations présentes dans le corpus. Cet ensemble contient plus de cent configurations différentes.

A partir des classificateurs les plus utilisés, nous obtenons un ensemble plus petit contenant onze configurations. Ces dernières couvrent près de la moitié des signes du corpus, car elles correspondent aux configurations les plus fréquemment rencontrées. Il s'agit des configurations **c**, **boule**, **0**, **plat**, **moufle**, **2**, **index**, **s**, **5**, **v** et **1**.

Ce sont ces configurations qui seront choisies en priorité pour élaborer les corpus durant la mise au point du système de reconnaissance.

Par ailleurs, nous avons pu constater que la réalisation articulatoire d'une configuration reste toujours à un niveau acceptable de complexité : il existe peu de configurations dynamiques et ces dernières sont constituées de configurations initiales et finales simples. De plus, l'annulaire et l'auriculaire ont un rôle passif dans les configurations dynamiques.

Ces informations seront utilisées lorsque l'on voudra intégrer aux corpus des gestes dont la configuration est dynamique.

2.3.2. MOUVEMENT

2.3.2.1. Base de données

Notations

Les mouvements sont spécifiés par la forme de la trajectoire du poignet dans l'espace. Ils sont décrits à l'aide de primitives géométriques telles que *droite*, *arc*, *cercle*. Pour stocker cette information, un champ nommé **Primitive Mouv** est utilisé

Afin de différencier tous les types de mouvement, leur trajectoire a été définie par rapport aux différents axes et plan d'un repère fictif placé sur le signeur (Figure 2.23).

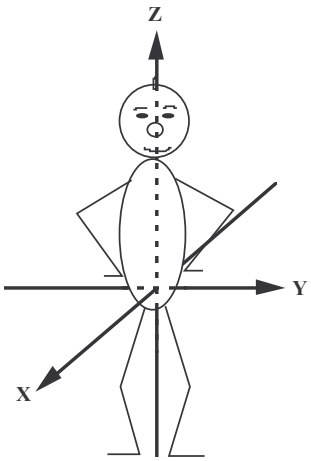


Figure 2.23 : Le repère "signeur".

Par la suite, la notation suivante est utilisée pour spécifier la trajectoire du mouvement dans l'espace.

- [X] : trajectoire parallèle à l'axe X (axe sagittal)
  [Y] : trajectoire parallèle à l'axe Y (axe latéral)
  [Z] : trajectoire parallèle à l'axe Z (axe vertical)
- [X,Y] : trajectoire parallèle au plan XY (plan horizontal)
  [Y,Z] : trajectoire parallèle au plan YZ (plan frontal)
  [X,Z] : trajectoire parallèle au plan XZ (plan sagittal)
- [V] : trajectoire dans un plan vertical quelconque
- [X,Y,Z] : trajectoire quelconque dans l'espace

Ainsi, par exemple, on pourra définir le mouvement du signe présenté Figure 2.24 en indiquant qu'il s'agit d'une droite de type [X].

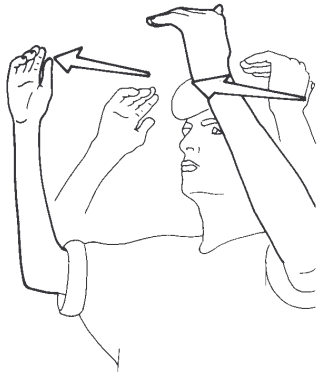


Figure 2.24 : Le signe [**plafond**].

Pour stocker cette information, un champ nommé **Sous-Type Mouv** est utilisé.

Pour chaque primitive, il faut aussi indiquer le sens du mouvement. Dans l'exemple précédent, le sens est positif par rapport à l'orientation de l'axe X.

Pour stocker cette information, un champ nommé **Sens Mouv** est utilisé.

Notons que sur l'axe [Y], le sens des signes s'inverse selon que le signeur est gaucher ou droitier. Pour simplifier les notations, le mouvement des signes est décrit en supposant qu'ils sont exécutés par un droitier.

Répétition

Il arrive qu'un mouvement du bras soit répété. Cela permet, par exemple, de différencier les noms des verbes. Ainsi, les signes [**boisson**] et [**boire**] se différencient par le fait que le mouvement est répété pour le nom (voir Figure 2.25).

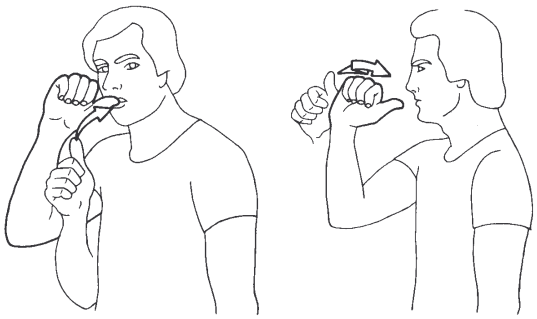


Figure 2.25 : Les signes [**boire**] et [**boisson**].

Il est possible aussi que le mouvement soit composé d'un aller et d'un retour, comme par exemple le mouvement du signe [**rideau**] (Figure 2.26).

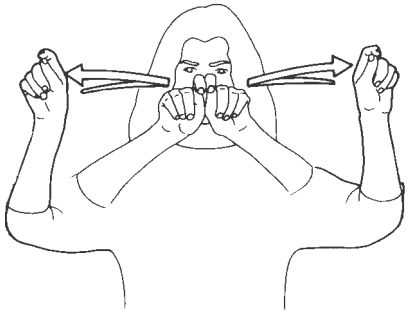


Figure 2.26 : Le signe [**rideau**].

La répétition éventuelle du mouvement devra être spécifiée, dans un champ nommé **Répèt Mouv**.

Rapport entre les mouvements des deux bras

Comme nous l'avons vu précédemment, certains signes sont réalisés avec une seule main, et d'autres avec les deux mains. Quand les deux mains entrent en jeu, plusieurs situations sont possibles :

- les deux mains font un mouvement différent (signe [**nouveau**], Figure 2.27, la main dominée est immobile tandis que la main dominante trace un arc vertical)

Figure 2.27 : Le signe [**nouveau**].

- les deux mains font un mouvement identique et de même sens (signe [**plafond**], Figure 2.24)

- les deux mains font un mouvement identique et de sens opposé (signe [**rideau**], Figure 2.26)
- les deux mains font un mouvement identique mais décalé dans le temps (signe [**échanger des idées**], Figure 2.28)

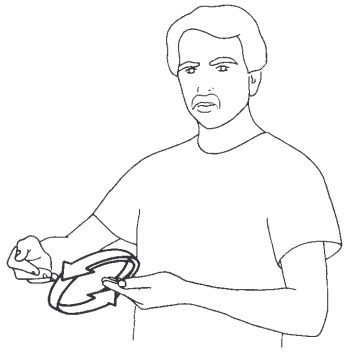


Figure 2.28 : Le signe [**échanger des idées**].

Ces rapports entre les mouvements des deux mains étant discriminants, ils devront eux aussi être spécifiés, dans un champ nommé **M1M2**.

Liste des champs

Finalement, pour le paramètre de mouvement, les cinq champs suivants ont été définis :

- **Primitive Mouv** : Primitive de mouvement de la main dominante
- **Sous-Type Mouv** : Sous-type de mouvement de la main dominante
- **Sens Mouv** : Sens du mouvement de la main dominante
- **Répèt Mouv** : Répétition éventuelle du mouvement de la main dominante
- **M1M2** : Rapport entre le mouvement des 2 mains, le cas échéant

Exemple

Un exemple de codage des mouvements est donné pour le signe [**rideau**] (Figure 2.26).

- **Primitive Mouv** : droite
- **Sous-Type Mouv** : [Y]
- **Sens Mouv** : négatif
- **Répèt Mouv** : aller-retour
- **M1M2** : opposé

2.3.2.2. Analyse

Ce paragraphe présente les résultats de l'analyse effectuée sur le paramètre de mouvement. Les primitives de mouvement sont peu nombreuses, mais selon le type de primitive, leurs "instanciations" dans l'espace peuvent être très nombreuses. Les principaux types de primitives ont été étudiés plus en détail, ainsi que les rapports entre les mouvements des deux mains.

Liste des primitives de mouvements

L'étude réalisée sur le dictionnaire de Bill Moody nous a permis de constater qu'il y avait très peu de primitives. Elles sont les suivantes :

- *statique :* le bras reste immobile ;
- *droite :* le bras trace une droite dans l'espace ;
- *arc :* le bras trace un arc dans l'espace ;
- *cercle* : le bras trace un cercle dans l'espace ;
- *complexe :* la trajectoire du bras dans l'espace est une trajectoire complexe (sinusoïde, zigzag ou autre).

Les proportions des différentes classes de mouvements sont présentées dans le tableau suivant.

Classe	%
droite	42,9
arc	26,1
statique	17,3
cercle	10,9
complexe	2,8

Tableau 2.1 : Proportions des différentes classes de mouvement.

Le principal résultat est que les mouvements les plus simples à réaliser sont aussi les plus fréquents, ce qui va dans le même sens que les observations concernant le paramètre de configuration. En particulier, les mouvements rectilignes sont présents dans plus de 40 % des signes.

Dans chacune de ces catégories, des sous-classifications sont possibles, en fonction de la direction du mouvement dans l'espace. Cependant, il existe une catégorie de signes pour

laquelle cette sous-classification est impossible. Il s'agit des verbes directionnels, présentés au Chapitre 1 (Paragraphe 1.4.2.4). Rappelons que la direction du mouvement du verbe permet de déterminer l'agent et le patient de l'action (exemple : [**je t'envoie**], Figure 1.19). De ce fait, la direction du mouvement et le paramètre d'orientation sont variables. Pour toutes les conjugaisons possibles de ces verbes, les seuls paramètres invariables sont la configuration et la primitive de mouvement. Notons que si un classificateur est combiné au verbe directionnel en tant que "super-pronom", seule la primitive de mouvement est invariable, quelle que soit la conjugaison du verbe.

La primitive statique

La primitive statique concerne les gestes pour lesquels le bras reste immobile. Cela ne représente que 17,5% des cas.

Il est intéressant de noter que ces gestes possèdent soit une configuration dynamique (signe [**éponge**], Figure 2.3), soit une variation de l'orientation, provoquée par une rotation (signe [**champagne**], Figure 2.29) ou une flexion du poignet. Les gestes dont tous les paramètres sont statiques sont extrêmement rares (dans le corpus étudié, seul le signe [**pipe**] est statique, mais dans ce cas c'est le cinquième paramètre, la mimique faciale, qui est dynamique : mouvement d'aspiration avec la bouche).

Figure 2.29 : Le signe [**champagne**].

A partir de ces remarques, on peut supposer qu'un comportement dynamique minimum doit être présent pour que l'existence du signe soit perceptible, mais que la complexité globale du signe doit rester à un niveau acceptable pour que la communication soit efficace.

La primitive droite

A l'aide du repère décrit au Paragraphe 2.3.2.1., nous avons étudié la répartition des différents signes de la classe droite dans les sous-catégories définies à l'aide du repère "signeur". Les pourcentages d'occurrence ont été calculés pour chaque sous-catégorie dans le dictionnaire, dans le but, comme pour le paramètre de configuration, de connaître le taux de couverture du système de reconnaissance en fonction des mouvements qu'il peut classifier. Ces taux sont donnés en Annexe 2.2.

Plus de 82 % des signes de la classe droite sont parallèles à un des 3 axes X, Y ou Z, près de 15 % sont parallèles à un des 3 plans formés par les axes et moins de 3 % sont dans un plan non orthogonal aux axes.

La primitive arc

La même étude a été réalisée pour la primitive arc. Près de 88 % des signes de type arc sont parallèles à un des 3 plans formés par les axes et environ 12 % sont dans un plan non orthogonal aux axes.

Les trois premiers types d'arcs ont été étudiés plus en détail. Pour différencier tous les sous-types possibles, nous avons dû concevoir une notation spécifique définie par le repère présenté Figure 2.30. Le sens de rotation est défini par les symboles + et -. Pour chaque sens de rotation, le plan est subdivisé en 8 zones. Selon le plan étudié, ces huit zones représentent des directions différentes dans l'espace.

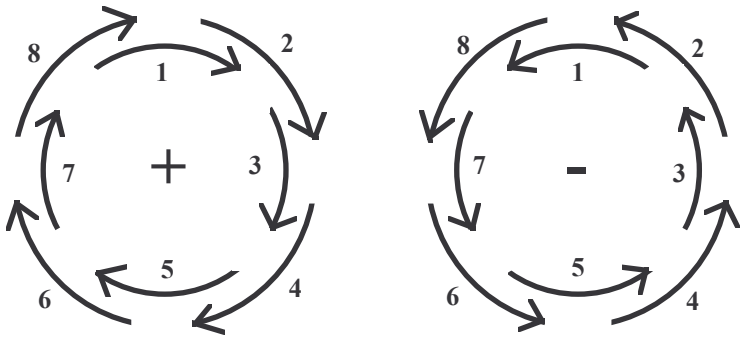


Figure 2.30 : repères pour les différentes sous-classes de la classe arc.

Pour les trois plans étudiés, les sens de rotations sont illustrés Figure 2.31 :

- Plan sagittal [X,Z] : le signeur est de profil dans la figure.

- Plan frontal [Y,Z] : le signeur est vu de dos dans la figure.
- Plan horizontal [X,Y] : le signeur est vu du dessus dans la figure.

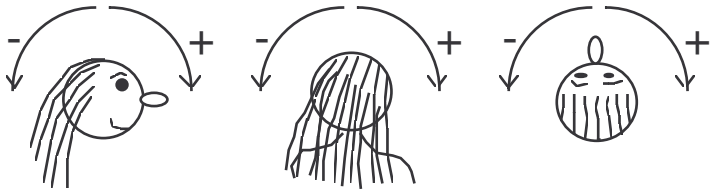


Figure 2.31 : Sens de rotations.

Par exemple, pour le plan sagittal, le symbole **+1** représente un arc vers l'avant, tandis que pour les plans frontal et horizontal, il représente un arc vers la droite.

Les résultats obtenus sont présentés en détail en Annexe 2.2. La Figure 2.32 illustre les principaux résultats. Les cas les plus fréquemment rencontrés sont indiqués à l'aide des arcs fléchés.

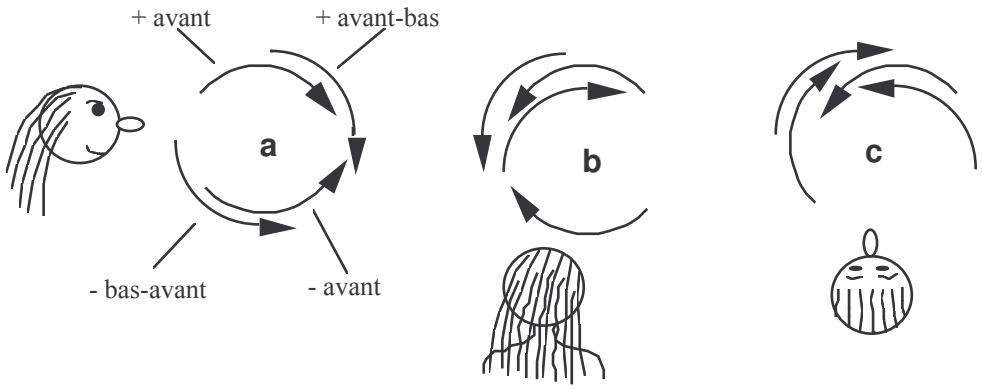


Figure 2.32 : Arcs les plus fréquents.

- Dans le plan sagittal [X,Z], les sous-classes les plus fréquentes sont +(avant-bas), +(avant), -(bas-avant) et -(avant) (Figure 2.32 - a).
- Dans le plan frontal [Y,Z], les sous-classes les plus fréquentes sont +(haut-droite), +(gauche), -(gauche-bas) et -(gauche) (Figure 2.32 - b).
- Dans le plan horizontal [X,Y], les sous-classes les plus fréquentes sont +(avant-droite), +(avant), -(gauche) et -(avant-gauche) (Figure 2.32 - c).

Une dissymétrie apparaît pour les plans frontal et horizontal sans doute due au fait que le signeur est droitier. Une dissymétrie inverse serait sans doute observable pour un gaucher.

La primitive cercle

Pour les signes ayant une primitive cercle, 92% sont parallèles à un des 3 plans formés par les axes et 8 % sont dans un plan non orthogonal aux axes.

Pour les signes dont la trajectoire est parallèle à un des trois plans, nous avons différencié les cercles effectués dans le sens des aiguilles d'une montre et ceux effectués en sens inverse, en nous plaçant selon les mêmes points de vue que pour les arcs.

Dans chacun des plans, un sens est plus souvent utilisé. Ces sens sont illustrés dans la Figure 2.33.

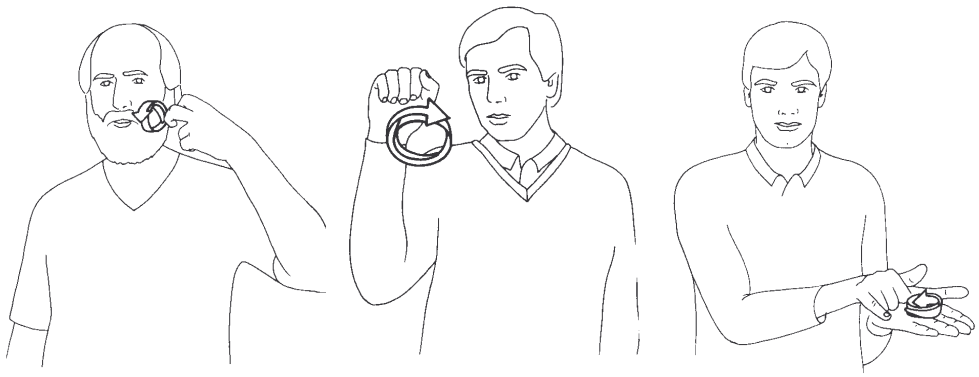


Figure 2.33 : Les signes [**train**], [**effacer**] et [**aéroport**].

Répétition

Certains types de mouvements sont plus souvent répétés que d'autres. Les mouvements de type arc ne sont pas souvent répétés, tandis que les mouvements de type statique et cercle sont souvent répétés. Quant aux mouvements de type droite, pour moitié ils sont répétés, pour moitié ils ne le sont pas.

Différences entre les mouvements des deux mains

Nous avons pu constater que les différences entre les mouvements des deux mains dépendent du type de mouvement (droite, arc, cercle) et surtout de son sous-type. Les associations les plus fréquentes sont indiquées dans le tableau ci-après.

Les symboles suivants ont été choisis :

- **1** : indique que seul le bras dominant est mobile ;
- **=** : indique que les deux bras sont mobiles et les mouvements sont parallèles ;
- **o** : indique que les deux bras sont mobiles et les deux mouvements sont opposés ;
- **dk** : indique que les deux bras sont mobiles et les deux mouvements sont décalés.

Lorsque deux symboles sont présents dans une case, le premier est plus fréquemment rencontré que le deuxième.

Primitive	[X]	[Y]	[Z]	[X,Y]	[Y,Z]	[X,Z]
Droite	= et 1	o et 1	1 et =	1	o	1
Arc				o et 1	o et 1	1 et =
Cercle				1	o	dk et 1

Tableau 2.2 : Rapports les plus fréquents entre les mouvements et le nombre de mains.

Les remarques suivantes sont déduite du tableau :

- Pour l'axe [Z] et les plans [X,Y] et [X,Z] (mouvements globalement verticaux), les signes pour lesquels un seul bras est mobile sont fréquemment rencontrés.
- Pour l'axe [Y] et le plan [Y,Z] (mouvements globalement latéraux), les signes pour lesquels deux bras sont mobiles et de mouvement opposé sont fréquemment rencontrés.
- Pour l'axe [X] (mouvements horizontaux, vers l'avant ou l'arrière), les signes pour lesquels deux bras sont mobiles et de mouvement parallèle sont fréquemment rencontrés.

De plus, de manière générale, quand les deux bras sont mobiles, ils ont toujours des mouvements symétriques (parallèles, opposés ou décalés) ; ils ne possèdent pas deux dynamiques indépendantes et difficiles à synchroniser.

Ces informations pourront être utiles si l'on désire par la suite développer un système de reconnaissance du mouvement des deux mains.

Dynamique

L'analyse effectuée n'apporte pas d'information concernant la dynamique des mouvements. Il est en effet difficile de détecter avec précision à partir d'une image si le mouvement est rapide, lent, ample… Cette information est cependant importante car elle permet de distinguer certains aspects d'un verbe. Elle intervient aussi dans l'expression de l'impératif. Elle ne pourra être étudiée que lorsque l'on disposera d'un véritable corpus sous forme dynamique.

Synthèse

L'étude réalisée sur le paramètre de mouvement nous a permis de constater que ce paramètre est basé sur des primitives géométriques simples : droite, arc et cercle. Les trois types de primitives peuvent être décomposées en sous-types définis par rapport au repère du signeur. Les primitives Droite et Cercle sont décomposables en peu de sous-types. La primitive arc est décomposable en de nombreux sous-types. Ces primitives s'exécutent le plus souvent le long d'une droite ou dans un plan parallèle aux axes et plans du repère du signeur.

Pour les verbes directionnels, la primitive de mouvement est parfois le seul paramètre à être invariant quelques soient leurs conjugaisons.

2.3.3. ORIENTATION

2.3.3.1. Base de données

Notations

La description de l'orientation de la main est basée sur le système utilisé par les concepteurs de HamNoSys [Prillwitz S. et Leven R. 1989], un système de notation dédié aux langues des signes et qui permet de coder à l'aide de symboles iconiques n'importe quel geste effectué par la main (voir Chapitre 1, Paragraphe 1.2.1.2.).

L'orientation de la main est définie par trois informations :

- l'état du poignet (plié ou tendu, au repos ou en rotation),
- la direction de l'axe qui passe par le dos de la main par rapport au repère du signeur,
- la direction du vecteur orthogonal au plan formé par la paume de la main par rapport au repère du signeur.

La Figure 2.34 montre les deux directions explicitant l'orientation de la main dans le cas où le poignet est tendu.

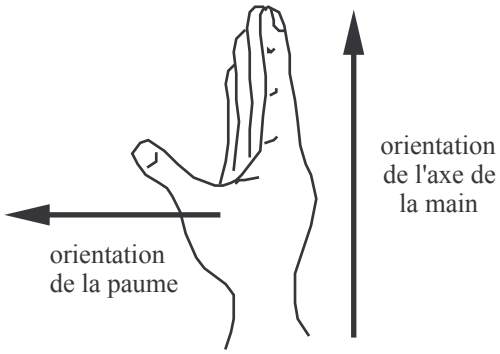


Figure 2.34: Définition de l'orientation de la main.

Comme nous l'avons vu précédemment, il est possible qu'un mouvement du poignet soit effectué durant un geste dont la primitive de mouvement est statique. Ceci est vrai aussi pour tous les autres types de mouvements, (droite, arc, cercle…). Du fait que l'orientation soit définie par deux directions mais aussi par la configuration du poignet, une modification de cette dernière entraîne une variation de l'orientation. Il est donc indispensable de préciser les modifications éventuelles de la configuration du poignet durant l'exécution du signe.

Liste des champs

Pour le paramètre d'orientation, les quatre champs suivants ont été définis :

- **Axe Main** : Direction de l'axe de la main
- **Paume** : Direction de la paume de la main
- **Poignet** : Configuration du poignet
- **Mouv Poignet** : Éventuel mouvement du poignet

Exemple

Exemple de codage de l'orientation sur le signe [**champagne**] (Figure 2.29) :

- **Axe Main** : + [Z]
- **Paume** : + [X] / - [X] (cette direction varie de + [X] à - [X])
- **Poignet** : tendu
- **Mouv Poignet** : rotation

2.3.3.2. Analyse

L'orientation est un paramètre très variable. Une étude sur les rapports entre le paramètre d'orientation et le paramètre de mouvement a été effectuée afin de mieux comprendre l'origine de cette grande variabilité. Cette dernière va poser beaucoup de problèmes pour un système de reconnaissance. Il est nécessaire d'étudier plus précisément le type d'information que véhicule l'orientation, afin d'en déduire son rôle fonctionnel.

Liste des orientations

Pour le paramètre d'orientation, plus de 300 valeurs différentes ont été dénombrées. Notons que les douze orientations de la main les plus fréquentes sont rencontrées dans plus de la moitié des signes du corpus. Les orientations les plus utilisées sont répertoriées en Annexe 2.3.

Liens entre orientation et mouvement

Selon le type de mouvement considéré, le comportement du paramètre d'orientation est différent. Pour les mouvements de type droite ou cercle, l'orientation est, la plupart du temps, statique. Par contre, pour les mouvements de type arc ou statique, l'orientation est plutôt dynamique.

Nous avons observé que les variations d'orientation sont en général dues à un mouvement du poignet. Ce mouvement peut être une rotation, une flexion, ou encore les deux simultanément.

De plus, on observe que dans les mouvements de type arc, dans certains cas, les variations d'orientation sont dues au mouvement du bras. Par exemple, dans la Figure 2.35, la direction de l'axe de la main varie durant le signe. Ce qui est significatif au niveau orientation dans ce signe, ce sont les directions des paumes des deux mains, statiques, qui délimitent les murs du couloir.

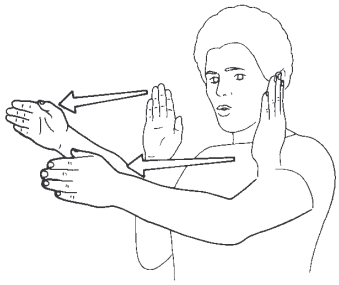


Figure 2.35 : Le signe [**couloir**].

Du fait du grand nombre d'orientations rencontrées dans le corpus d'analyse, il semble difficile d'en tirer une liste minimum pouvant permettre de mettre au point un corpus de reconnaissance.

Cette diversité est certainement due au fait que pour une grande partie des signes étudiés, l'orientation de la main semble être une conséquence du geste effectué, plutôt qu'une orientation intentionnellement choisie par rapport au repère "signeur". C'est le cas pour la direction de l'axe de la main dans le signe [**couloir**].

Informations véhiculées

Les signes pour lesquels l'orientation est volontaire sont, en général, ceux qui servent à décrire des informations de type spatial. Dans ce cas, le signeur cherche à indiquer l'orientation d'un objet ou d'une personne dans la scène du signeur. Elle est alors choisie par rapport au repère "signeur". Cette information orientographique est nécessaire pour interpréter les signes non standard, tels que les classificateurs, mais aussi les verbes directionnels présentés au Chapitre 1, ainsi que les gestes déictiques, qui sont employés, par exemple, en association avec les verbes non directionnels.

Le dessin d'une forme dans l'espace, lorsqu'il est dynamique, fait intervenir la configuration, le mouvement, mais aussi l'orientation. Par exemple, dans le signe [**armoire**] (Figure 2.36), le meuble est "dessiné" dans l'espace. Pour "tracer" le dessus d'un meuble, le signeur dirige la paume vers le bas, tandis que pour le côté, la paume est dirigée vers la gauche.

Figure 2.36 : Le signe [**armoire**][2].

Ce signe est un signe standard, mais il n'est pas arbitraire et il pourrait être utilisé dans certains contextes pour représenter d'autres objets ayant la même forme, tels qu'un immeuble, un carton, une machine…

Synthèse

L'orientation, ou plutôt la variation d'orientation dans le cas d'une orientation dynamique, peut être corrélée au mouvement du bras. Ce cas se rencontre souvent dans les mouvements de type arc.

Le paramètre d'orientation possède un statut différent des paramètres vus précédemment. Comme les deux premiers, il sert à différencier les signes entre eux. Cependant, dans le cas où l'orientation est choisie délibérément, il nécessaire de connaître sa valeur précise par rapport au repère du signeur pour pouvoir interpréter le signe.

En conclusion, le contenu sémantique du paramètre d'orientation a une importance variable et il est nécessaire de connaître la fonction syntaxique du signe pour en déduire s'il possède une valeur sémantique.

Remarque :

Il serait sans doute intéressant de connaître les valeurs d'orientation de la main par rapport à un repère relatif placé sur l'avant-bras du signeur. Ainsi, le nombre de valeurs

2 Le signe armoire a été considéré comme étant un signe composé car il décrit la forme de l'objet à l'aide de deux segments, un pour le dessus et un pour les côtés. Il n'a donc pas été inclus dans la base de données.

d'orientation possibles serait considérablement réduit. De plus, lorsqu'une orientation dynamique apparaîtrait, cela correspondrait réellement uniquement au cas où un mouvement du poignet est présent. Cependant, l'information véhiculée par l'orientation, lorsqu'il y en a une, est toujours interprétable relativement au repère du signeur. Les valeurs d'orientation telles que définies précédemment doivent donc être gardées.

2.3.4. EMPLACEMENT

2.3.4.1. Base de données

Il existe un nombre fini d'emplacements génériques relatifs au signeur.

Sur le corps, les quinze localisations principales sont :

- la tête, le front, les yeux, l'oreille, le nez, la bouche, la joue, le menton, le cou,
- le torse, le coeur, la partie droite du torse, le ventre,
- le coude, le bras.

Dans l'espace, les trois zones principales sont :

- la zone située devant le buste,
- la zone située devant la tête,
- la zone située à coté de la tête.

Certains signes se font dans une seule zone, d'autres passent d'une zone à l'autre durant le signe.

Pour le paramètre d'emplacement, un seul champ informatif nommé **Emplacement** est défini. Il précise la zone ou les zones en cas d'emplacement variable.

2.3.4.2. Analyse

L'emplacement est un paramètre dont le rôle est variable. Il est nécessaire d'étudier plus précisément le type d'information que véhicule l'emplacement, afin d'en déduire son rôle fonctionnel.

Liste des emplacements

Les emplacements les plus fréquents sont répertoriés en Annexe 2.4. En tout, 48 emplacement différents ont été identifiés, les 14 premiers correspondant à eux seuls à 92,6% des signes.

Les emplacements correspondent à des zones dans l'espace. Bien que le mouvement du bras provoque un changement d'emplacement, dans 95,8% des cas, les emplacements sont considérés comme étant statiques au sein de ces zones.

Pour une grande partie des signes étudiés (plus de 60%), l'emplacement de la main se situe dans la zone neutre placée dans une demi-sphère face au signeur. Dans ce cas, l'emplacement n'a pas obligatoirement de fonction syntaxique ou de valeur sémantique particulière (exemple, Figure 2.37).

Figure 2.37 : Le signe [**quoi ?**].

On remarque que les verbes qui font référence à une fonctionnalité d'une partie du corps s'effectuent en général dans cette zone du corps. Ainsi, par exemple, tous les signes qui traitent d'une activité cérébrale (Figure 2.38) sont en général exécutés dans une zone proche du front ou de la tête.

Figure 2.38 : Les signes [**comprendre**], [**intelligent**] et [**rêver**].

Pour ces signes, une connaissance grossière de la zone dans laquelle le signe a été effectué est suffisante pour reconnaître le signe.

Informations véhiculées

Mais pour d'autres signes, l'emplacement est volontairement choisi. C'est le cas des verbes dans lesquels l'emplacement permet d'incorporer une partie du corps comme nous l'avons indiqué dans le Chapitre 1 (signe [**opérer**]). Dans ce cas, une connaissance un peu plus précise de l'emplacement est nécessaire pour interpréter le signe.

C'est aussi le cas des signes qui servent à décrire des informations de type spatial. Pour décrire une situation, une action, le signeur commence par placer le contexte : il associe à chaque personne ou objet intervenant dans l'histoire une place dans une scène fictive placée devant lui.

Par exemple, dans la Figure 2.39, le signeur place un garçon à sa droite. Un classificateur est utilisé pour indiquer l'emplacement du garçon. Ensuite, il pourra indiquer cet emplacement à l'aide d'un geste déictique lorsqu'il voudra se référer au garçon.

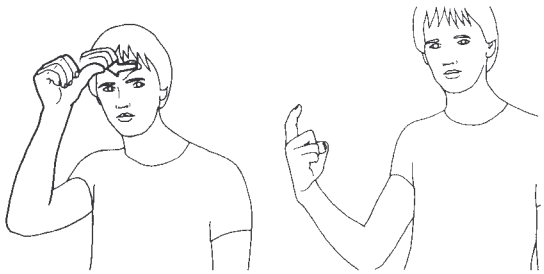


Figure 2.39 : Phrase "**un garçon à droite**".

L'emplacement permet de localiser des personnes, des objets ou des événements sur la scène du signeur et de montrer les relations spatiales entre les entités. Selon la classification définie dans le Chapitre 1, il s'agit d'informations spatiographiques.

S'il s'agit d'un classificateur, d'un déictique ou d'un verbe directionnel, alors il faut récupérer la valeur de l'emplacement, afin de pouvoir interpréter la totalité du sens du signe. S'il s'agit d'un simple substantif, il n'est pas nécessaire de connaître précisément la valeur de l'emplacement. Cette dernière permet tout au plus de différencier deux signes dont tous les autres paramètres sont identiques.

Liens entre emplacement et autres paramètres

Les configurations et les mouvements les plus complexes sont effectués près du visage. Tandis que si les signes sont effectués loin du visage, ils ont tendance à être symétriques, avec les deux mains, avec des configurations simples ou des mouvements plus grands.

Ceci est dû au fait que l'interlocuteur centre son regard sur le visage du signeur et les signes éloignés du visage sont perçus par sa vision périphérique. C'est sur le visage qu'il a la plus grande précision visuelle [Moody B. 1983].

Synthèse

L'emplacement est corrélé au mouvement du bras.

Le paramètre d'emplacement possède un statut assez semblable à celui d'orientation. Il sert à différencier les signes entre eux, mais de plus, dans le cas où l'emplacement est choisi délibérément, il est nécessaire de connaître sa valeur précise par rapport au repère du signeur pour pouvoir interpréter le signe.

En conclusion, de même que pour le paramètre d'orientation, le contenu sémantique du paramètre d'emplacement à une importance variable et il est nécessaire de connaître la fonction syntaxique du signe pour en déduire s'il possède une valeur sémantique.

2.4. AUTRES INFORMATIONS

Ce paragraphe contient la liste des champs contenant des informations d'ordre général, tel que le nom du signe. Un champ permettant de stocker des informations sur les contacts éventuels entre mains et corps durant le signe y a été adjoint.

2.4.1. BASE DE DONNEES

2.4.1.1. Notations

Il peut arriver que l'émission d'un signe soit accompagnée de contacts entre les deux mains ou entre la main et une partie du corps. Deux types d'information doivent être spécifiés en cas de contact.

1. Le moment durant lequel le contact a lieu :
   - *aucun* : aucun contact n'est présent durant le signe,
   - *fin* : contact en fin de signe,
   - *début* : contact en début de signe,
   - *milieu* : contact en milieu de signe,
   - *toujours* : contact pendant toute la durée du signe,
   - *début + fin* : contact en début et fin de signe.
2. Les parties du corps concernées :
   - *2 mains* : contact entre main dominante et main dominée,
   - *main - bras* : contact entre main dominante et bras opposé,
   - *main - corps* : contact entre main dominante et une partie du corps,
   - *main - tête* : contact entre main dominante et une partie de la tête.

Ces contacts peuvent être produits dans le cadre d'une configuration dynamique, ou dans le cadre du paramètre mouvement. Cette information a été dissociée des informations relatives aux différents paramètres. Elle est stockée dans un champ nommé **Contact**.

2.4.1.2. Liste des champs

Les informations dites d'ordre général sont les suivantes :

- **Num** : le numéro du signe dans le dictionnaire de Moody
- **Traduc** : le sens du signe en français
- **Nb Mains** : le nombre de mains utilisées
- **Contact** : la présence éventuelle d'un contact durant le signe
- **Remarques** : variantes, synonymes

2.4.2. ANALYSE

En ce qui concerne les contacts, nous avons constaté que dans plus de la moitié des cas, un contact se produit durant le signe, dont plus de la moitié entre les deux mains et plus du quart entre les mains et le visage (Annexe 2.5). Le reste est essentiellement composé de contact entre les mains et le corps et quelques rares contacts entre la main et le bras opposé.

Ces informations pourront être utiles si par la suite on souhaite adjoindre au système de capture de geste un outil permettant de détecter les contacts.

Bien d'autres résultats peuvent être obtenus à partir de cette base de données. Par exemple, il est possible d'observer la corrélation entre la complexité du paramètre de configuration (statique, dynamique) et la complexité du paramètre de mouvement (statique, droite, arc, cercle, une main, deux mains…). Comme l'on pouvait s'y attendre, plus le mouvement est compliqué, plus la configuration est simple et réciproquement.

2.5. LA BASE DE DONNEES

Dans ce paragraphe, nous rappelons tout d'abord l'ensemble des champs utilisés dans la base de données, avec pour chaque champ la liste des choix possibles donné entre accolades ou un renvoi aux annexes quand la liste est trop longue. Puis un exemple de description d'un signe, le signe [**immeuble**], est présenté suivant les différents champs.

La base de données est disponible sur demande pour toute personne intéressée (format Paradox, sur PC).

2.5.1. STRUCTURE COMPLETE

La base de données comporte l'ensemble des champs suivants :

Informations générales sur le signe :

- **Num** : Numéro du geste (dans le premier tome du dictionnaire de 1 à 1359)
- **Traduc** : Signification du geste (traduction du dictionnaire)
- **Nb Mains** : Nombre de mains mises en jeu {1, 2}
- **Contact** : Présence éventuelle d'un contact durant le signe.
  temps : {aucun, fin, début, milieu, toujours, début+fin}
  parties du corps : {2mains, main/bras, main/corps, main/tête}
- **Remarques** : Remarques diverses (synonymes, variantes)

Informations sur le paramètre de configuration :

- **Config 1** : Configuration de la main dominante (liste en Annexe 2.1)
- **Config 2** : Configuration de la main dominée (liste en Annexe 2.1)
- **Type Config** : Type de comportement dynamique de la main dominante {statique, fermeture, ouverture, vibration, frottement, fermeture/ouverture, ouverture/fermeture, ouverture/vibration}
- **Répèt Config** : Répétition éventuelle en cas de configuration dynamique de la main dominante {oui, non}
- **C1C2** : Rapport entre la configuration des 2 mains le cas échéant {identiques, différentes}

Informations sur le paramètre de mouvement :

- **Primitive Mouv** : Primitive de mouvement de la main dominante {statique, droite, arc, cercle, complexe}

- **Sous-Type Mouv** : Sous-type de mouvement de la main dominante
  {[X], [Y], [Z], [X,Y], [X,Z], [Y,Z], [V], [X,Y,Z], 1, 2, 3, 4, 5, 6, 7, 8}
- **Sens Mouv** : Sens du mouvement de la main dominante {+, -}
- **Répèt Mouv** : Répétition éventuelle du mouvement de la main dominante
  {oui, non, aller-retour}
- **M1M2** : Rapport entre le mouvement des 2 mains le cas échéant {1, =, o, dk}

Informations sur le paramètre d'orientation :

- **Axe Main** : Direction de l'axe de la main
  {±[X], ±[Y], ±[Z], ±[X,Y], ±[X,Z], ±[Y,Z], et les combinaisons de deux ou trois de ces directions (ex: +[X]/-[X]}
- **Paume** : Direction de la paume de la main
  {±[X], ±[Y], ±[Z], ±[X,Y], ±[X,Z], ±[Y,Z], et les combinaisons de deux de ces directions (ex: +[X]/-[X]}
- **Poignet** : Configuration du poignet {tendu, plié, rotation}
- **Mouv Poignet** : Éventuel mouvement du poignet {non, rotation, flexion}

Informations sur le paramètre d'emplacement :

- **Emplacement** : Emplacement du signe par rapport au corps du signeur
  (liste en Annexe 2.4)

2.5.2. EXEMPLE DE DESCRIPTION

Un exemple de description complète du signe [**immeuble**] est donné ici (Figure 2.40).

Figure 2.40 : Le signe [**immeuble**].

- **Num** : 3
- **Traduc** : immeuble
- **Nb Mains** : 2
- **Contact** : non
- **Remarques** : non
- **Config 1** : sc
- **Config 2** : sc
- **Type Config** : statique
- **Répèt Config** : non
- **C1C2** : identique
- **Primitive Mouv** : droite
- **Sous-Type Mouv** : [Z]
- **Sens Mouv** : +
- **Répèt Mouv** : non
- **M1M2** : identique
- **Axe Main** : + [X]
- **Paume** : + [Y]
- **Poignet** : tendu
- **Mouv Poignet** : non
- **Emplacement** : devant buste

2.6. CONCLUSION

Dans ce chapitre, une étude détaillée des quatre premiers paramètres de la LSF, ceux qui se rapportent à la main, a été présentée. Les résultats obtenus vont permettre de définir un système de reconnaissance et de compréhension de gestes et de choisir un corpus approprié afin de tester ce système. Les principaux résultats sont résumés ci-dessous.

Du point de vue reconnaissance, deux catégories de signes doivent être distingués : ceux pour lesquels les quatre paramètres sont invariables quelque soit le contexte et ceux pour lesquels au moins un des paramètres est variable en fonction du contexte. La première catégorie correspond aux signes standard. La seconde inclut les classificateurs et les verbes directionnels. Pour les classificateurs, seule la configuration est invariable. Pour les verbes directionnels, au sein du paramètre de mouvement, seule la primitive de la trajectoire est invariable : son "instanciation" dans l'espace varie en fonction de sa conjugaison. Pour ces verbes, la configuration peut intégrer un classificateur.

Du point de vue compréhension, deux catégories de paramètres doivent être distinguées selon que ces derniers possèdent une valeur sémantique ou pas. Les paramètres Configuration et Mouvement possèdent toujours une valeur sémantique. Ce n'est pas le cas des paramètres Orientation et Emplacement. Si l'orientation n'est pas choisie délibérément mais est une conséquence du mouvement, elle ne possède pas de valeur sémantique. Pour l'emplacement, cela se produit lorsque le geste est réalisé dans la zone neutre. En revanche, lorsque ces deux paramètres sont porteurs d'informations, leurs valeurs numériques précises sont nécessaires pour compléter l'interprétation.

Remarques :

Lors de l'analyse du corpus, nous avons tenté d'acquérir le plus d'informations possibles, même si celles-ci ne sont pas toutes exploitées dans le cadre de cette thèse. Elles serviront de base à diverses extensions possibles présentées dans les perspectives (Chapitre 5).

Une telle étude ne permet pas d'obtenir des données sur la fréquence d'utilisation des signes dans une conversation réelle entre signeurs. Des corpus en situation réelle sont difficiles à obtenir et leur dépouillement nécessite une très bonne connaissance de la LSF. Il faut pouvoir travailler en collaboration avec des personnes dont la langue maternelle est la LSF. L'étude d'un corpus constitué d'enregistrements de conversations réelles entre signeurs devra être réalisée dans une étape future afin de compléter les informations obtenues ici.

Chapitre 3

LA RECONNAISSANCE DE GESTES

L'objectif de ce chapitre est de proposer un système de reconnaissance de phrases gestuelles de la LSF. Ce système de reconnaissance doit prendre en compte les remarques exprimées dans le chapitre précédent, à savoir que deux catégories de signes doivent être distingués : ceux pour lesquels les quatre paramètres sont invariables quelque soit le contexte et ceux pour lesquels au moins un des paramètres est variable en fonction du contexte. Une étude des différents systèmes existants a été effectuée afin de choisir la technique la mieux adaptée à notre objectif qui impose le traitement de gestes dynamiques continus. Notons dès à présent que la technique de reconnaissance utilisée est fondée sur les modèles de Markov cachés qui ont permis l'obtention de taux de reconnaissance très encourageants, sur les deux catégories de signes.

Ce chapitre comporte trois parties principales. La première est un état de l'art dans le domaine de la reconnaissance de gestes. La deuxième présente d'une part le capteur utilisé, le gant numérique et d'autre part les différents outils qu'il a fallu développer afin de mettre au point le système de reconnaissance (affichages, filtrages, calibration, saisie des corpus, segmentation des corpus d'apprentissage…). La troisième partie présente l'architecture du système de reconnaissance, le corpus choisi et les premières évaluations de ce système.

3.1. ÉTAT DE L'ART

Les études réalisées dans le domaine de la reconnaissance de gestes sont récentes, du fait que les gants numériques n'ont fait leur apparition qu'en 1987, avec le DataGlove de VPL. Si depuis deux ou trois ans beaucoup d'études ont été réalisées, il s'agit pour la plupart d'études de faisabilité permettant de tester, pour un système de reconnaissance donné utilisé dans un domaine particulier (tel que la parole ou le geste 2D), s'il peut être utilisé dans le cadre du geste 3D.

C'est sans doute une des raisons pour lesquelles il ne s'est pas dégagé à ce jour un consensus en ce qui concerne le type de méthode à utiliser, de même qu'il n'existe pas de corpus, national ou international, permettant de comparer les performances des différentes méthodes.

Ces méthodes sont très variées, de même que le vocabulaire étudié : il peut s'agir de postures de la main, statiques, ou de gestes dynamiques. De plus, certaines études s'attachent à la reconnaissance de gestes isolés, tandis que d'autres s'intéressent à des séquences de gestes enchaînés. Dans ce dernier cas, différentes méthodes sont utilisées pour segmenter les gestes.

Après avoir présenté la terminologie relative au domaine de la reconnaissance de gestes, nous répertorions les méthodes les plus courantes dans le domaine de la reconnaissance de gestes 3D pour segmenter les données, les représenter et les classifier. Les applications les plus courantes sont ensuite décrites avec des tableaux comparatifs portant sur la taille du vocabulaire et les performances des systèmes. En conclusion, nous indiquons quels sont les choix qui nous semblent les meilleurs en fonction du type d'application choisi.

3.1.1. TERMINOLOGIE

La terminologie employée dans notre contexte est issue des domaines de la reconnaissance des formes en général et des gestes en particulier.

3.1.1.1. Reconnaissance des formes

La reconnaissance des formes peut se définir comme l'ensemble des techniques informatiques de représentation et de décision permettant aux machines d'interpréter des événements issus de capteurs physiques. L'interprétation consiste à catégoriser le phénomène perçu : il s'agit de passer d'une représentation numérique, c'est-à-dire continue, à une représentation symbolique, ou encore discrète. Il faut construire des programmes qui, à partir

de données topologiques à valeur dans un **espace de représentation** (X), permettent de décider automatiquement à quelle classe, dans un **espace d'interprétation** (Ω), appartient les données [Simon J. C. 1984]. Les systèmes de reconnaissance sont composé de deux sous-systèmes dédiés respectivement aux processus de **Représentation** et de **Décision** (Figure 3.1).

Figure 3.1 : Les deux étapes de la reconnaissance des formes.

- *Le processus de Représentation transforme les données brutes issues des capteurs en une représentation particulière de la forme.*

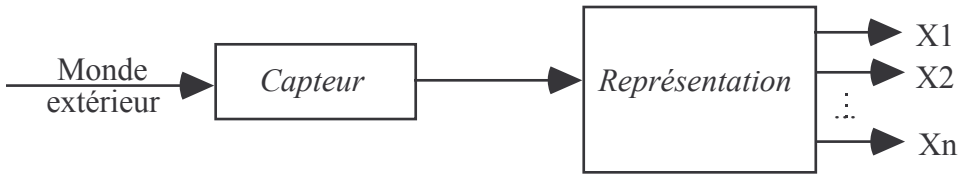


Figure 3.2 : Représentation d'une forme.

  La représentation interne de la forme est alors, en général, constituée d'un **vecteur de paramètres** extraits des données brutes (Figure 3.2). Notons que le mot *paramètre* possède ici un sens différent de celui employé au chapitre précédent concernant la Langue des Signes (configuration, mouvement, orientation, emplacement et expression faciale). Afin d'éviter toute confusion par la suite, le terme **paramètres de représentation** sera utilisé quand il sera question d'algorithmes de reconnaissance de forme.

- *Le processus de Décision prend en entrée la sortie du processus de Représentation et produit en sortie, si c'est possible, une classification de la forme.*

Les systèmes de reconnaissance travaillent en général sur un vocabulaire bien défini. Ce vocabulaire peut être appris par certains types de systèmes de reconnaissance. Cette étape est appelée **apprentissage** (Figure 3.3). Pour chaque unité de ce vocabulaire, ou **classe**, un **vecteur de référence** doit être déterminé. Le processus de Décision est chargé de comparer le vecteur de paramètres de représentation de la forme à reconnaître avec les différents vecteurs de référence et choisir celui qui est le plus proche.

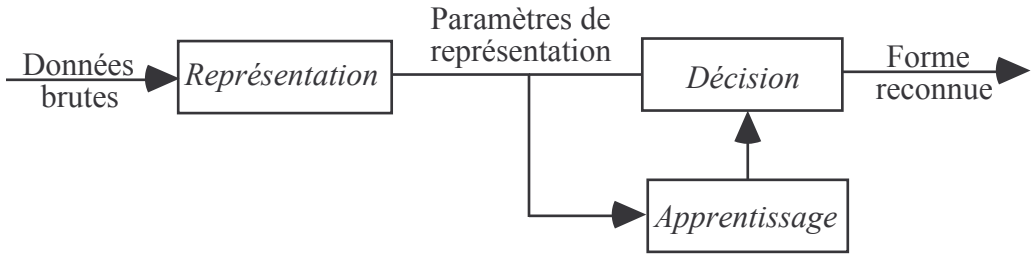


Figure 3.3 : Système de reconnaissance avec apprentissage.

3.1.1.2. Reconnaissance de gestes

Jusqu'alors, le terme geste a été utilisé pour représenter une information quelconque portée par une partie du corps en mouvement. Dans le domaine de la reconnaissance des formes, les approches utilisées dépendent du type de geste étudié. La forme de base peut être soit statique, soit dynamique. La terminologie employée ici est la suivante :

- Une **posture** correspond à un vecteur de données à un instant **t**. Il peut être constitué d'une configuration, d'un emplacement et d'une orientation de la main.
- Un **geste** est une séquence de postures.
- Les postures ou les gestes peuvent être étudiés isolément. On parle alors de postures ou de gestes **isolés**.
- Des postures ou des gestes sont **connectés** s'ils sont exécutés les uns à la suite des autres mais sans qu'il y ait de recouvrement entre les gestes. Cela peut se faire en choisissant une configuration et un emplacement arbitraire qui pourront être considérés comme l'équivalent du silence en parole.
- Des gestes sont **enchaînés** si la fin d'un geste est modifiée en fonction du début du geste suivant et si le début du geste suivant est modifié en fonction de la fin du geste précédent. Il s'agit du phénomène de **coarticulation**.

Le phénomène de coarticulation complique la reconnaissance. Considérons par exemple le cas suivant : le signe **b** est suivi du signe **o**. Pour passer de la première à la deuxième configuration, la lettre **c** va être formée durant un court instant (Figure 3.4). Il faut cependant que le système de reconnaissance reconnaisse la séquence **bo** et non pas **bco**.

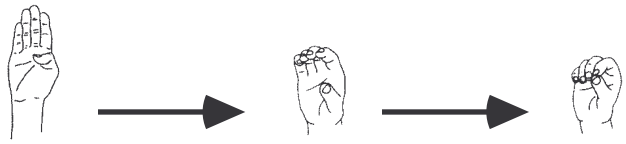


Figure 3.4 : Passage de **b** à **o** en passant par **c.**

Les autres difficultés potentielles sont les suivantes :

- *Variabilité du signal dans le temps*. Il existe à la fois des variations inter-personnes et intra-personnes dans la durée d'exécution d'un geste et dans la dynamique.
- *Variabilité du signal dans l'espace*. Si l'on considère tous les degrés de liberté existant dans le système articulé poignet/main sans même compter les articulations des métacarpiens dans la paume, on obtient un espace de données de 24 dimensions. Pour chacune de ces dimensions, il existe une variabilité du signal, même si certaines dimensions sont corrélées.
- Il n'existe pas d'indication explicite du *début et de la fin d'un geste*, contrairement à un geste réalisé par exemple avec une souris pour lequel peuvent être utilisés les boutons pour spécifier l'étendue du geste. Il faut concevoir un système permettant la **segmentation** automatique des gestes lorsqu'ils sont connectés ou enchaînés.
- *Informations multiples*. Comme nous l'avons vu dans les deux premiers chapitres, les gestes portent plusieurs informations en parallèle, par l'intermédiaire du mouvement, mais aussi de la configuration, de l'orientation et de l'emplacement.

Mise à part cette dernière difficulté, inhérente au geste de la main, les autres problèmes sont communs à toutes les formes de données dynamiques, comme la parole continue. Les techniques utilisées sont très variées et si elles prennent en compte une ou plusieurs des difficultés énumérées ci-dessus, rares sont celles qui permettent de toutes les gérer.

3.1.2. SEGMENTATION

La plupart des études concernant la reconnaissance de postures ou de gestes *connectés* ou *enchaînés* sont basées sur des outils permettant de reconnaître des postures ou des gestes *isolés*. Pour cela, les signaux gestuels sont segmentés par détection de caractéristiques telles que des pauses, des postures ou des points de rebroussement. Cette détection est faite préalablement à la reconnaissance. Par ce procédé, les séquences gestuelles sont décomposées en gestes segmentés, donc isolés. Il est alors possible d'utiliser des algorithmes de reconnaissance de gestes isolés, plus simples à mettre en oeuvre.

Cependant, les performances du système de reconnaissance risquent d'être limitées par les performances du système de segmentation. En effet, la détection des caractéristiques utilisées pour la segmentation est en général réalisée par comparaison avec des valeurs seuil fixées a priori qui risquent de ne pas être robustes lorsque le vocabulaire augmente ou que le nombre d'utilisateur augmente.

D'autres outils permettent d'effectuer de manière simultanée la reconnaissance et la segmentation d'une séquence de gestes. Dans le domaine du geste, les techniques les plus fréquemment rencontrées sont la comparaison dynamique et les modèles de Markov cachés (HMM).

Les différentes techniques sont présentées ci-après.

Détection de pauses

Une pause correspond à un geste pour lequel toutes les valeurs sont stables durant un certain temps (une posture). La détection de pause se fait simplement en calculant la variation des différentes valeurs du vecteur de données et en le comparant à des seuils $\mathbf{s_i}$. La taille **n** de la fenêtre temporelle d'observation est, elle aussi, définie a priori. Les études utilisant cette technique sont nombreuses [Cagin J.-M. 1993], [Da Silva Faria O. 1991], [Takahashi T. et Kishino F. 1991].

De manière à permettre la reconnaissance en continu sans faire de pauses artificielles entre chaque geste, une fenêtre d'observation est utilisée. Cette fenêtre contient les derniers gestes reconnus. Si un même geste est reconnu sur un nombre **m** prédéfini de trames, le geste est validé [Sturman D. J. 1992]. Ce système est utilisé en particulier pour résoudre le problème de coarticulation (exemple précédent **bo/bco**) [Liang R.-H. et Ming O. 1995].

Détection de configurations

La technique de détection de configuration est fondée sur la supposition que chaque geste commence et se terminent par des configurations bien définies. La segmentation consiste à vérifier si les configurations correspondent à une configuration de début ou de fin de geste [Murakami K. et Taguchi H. 1991] (début uniquement), [Braffort A. 1992], [Bordegoni M. et Hemmje M. 1993].

Autres techniques

La détection de points caractéristiques (valeurs cinématique - ex : vitesse nulle - ou géométrique - ex : point de rebroussement) peut servir à segmenter le mouvement [Sagawa H., Sakou H. et al. 1992], [Bellalem N. 1995]. Cette détection est réalisée à l'aide d'un apprentissage dans [Fels S. S. et Hinton G. E. 1993].

La détection de la fin d'un geste peut être détectée dans les réseaux connexionnistes par l'intermédiaire d'un seuil auquel est comparée l'activation des neurones de sortie. Si le niveau d'activation est supérieur à ce seuil, la fin du geste est détectée [Murakami K. et Taguchi H. 1991].

Segmentation simultanée

Certains outils tels que la comparaison dynamique [Cagin J.-M. 1993] et les HMM [Starner T. E. 1995] permettent une segmentation automatique qui est réalisée en même temps que la reconnaissance se fait. Ce type de segmentation ne nécessite pas de comparaison avec des seuils fixés a priori. Ce sont des techniques plus robustes. Elles sont décrites plus précisément dans le paragraphe dédié aux différentes techniques de décision.

Les seules comparaisons que l'on peut faire sur les différentes méthodes de segmentation sont d'ordre qualitatif. Les performances des systèmes de reconnaissance sont souvent mesurées de manière globale, sans distinction entre la partie segmentation et la partie classification. Le choix d'une technique de classification ou d'un vocabulaire spécifique induit souvent la méthode de segmentation utilisée. Les différentes techniques de représentation des données puis de décision sur ces données sont présentées dans les deux paragraphes suivants.

3.1.3. LA REPRESENTATION

Pour mettre en oeuvre une représentation pertinente, robuste et simple, il faut choisir les paramètres de représentation possédant certaines qualités qui sont la continuité, la sensibilité et réversibilité, l'indépendance des paramètres et l'homogénéité dans le temps [Geoffrois E. 1995] :

- Continuité

  Des signaux perceptiblement proches doivent être représentés par des paramètres similaires. Une faible variation du signal doit engendrer une faible variation de la valeur des paramètres. Pour respecter cette propriété, les caractéristiques qui

dépendent de seuils doivent être évitées. De manière idéale, un paramètre doit être une fonction réelle continue de l'entrée.

- Sensibilité et réversibilité

  Des signaux perceptiblement différents doivent être représentés par des paramètres discriminants, afin que l'information pertinente ne soit pas perdue dans les traitements. Si cette propriété est respectée, il est théoriquement possible de reconstruire un signal perceptiblement comparable au signal d'origine à partir des paramètres.

- Indépendance des paramètres

  Les corrélations entre les paramètres doivent être supprimées, afin d'éliminer l'information non pertinente.

- Homogénéité dans le temps

  Les paramètres doivent être disponibles à chaque trame et le traitement doit être le même pour toutes les trames.

A ces qualités s'ajoutent des critères de simplicité et rapidité des algorithmes de traitements du signal.

Les méthodes de représentation peuvent être classés selon le type de *paramètres de représentation* qu'ils fournissent. Les trois méthodes les plus répandues en reconnaissance de gestes sont les **prototypes** (template), la discrétisation de l'espace en **zones** et les **caractéristiques** cinématiques et géométriques. Par ailleurs, la construction des paramètres de représentation peut être combinée à une **approche statistique**.

3.1.3.1. Les prototypes

L'élaboration des prototypes ne nécessite aucun calcul : il s'agit simplement des données captées en entrée sous leur forme brute. Dans le cas du gant numérique, s'il est question de posture, les prototypes sont des vecteurs constitués des 10 valeurs de flexion des doigts [Cagin J.-M. 1993], [Da Silva Faria O. 1991], [Gourley C. 1994], [Messing L. S., Erenshteyn R. et al. 1994] et [Newby G. B. 1993]. Parfois, les trois valeurs d'orientation sont également utilisées [Murakami K. et Taguchi H. 1991]. S'il s'agit de gestes dynamiques, les prototypes sont constitués d'une séquence complète de postures (doigts, orientation, position) d'une taille fixe [Collet C. 1993], [Harling P. A. 1993].

Le défaut des prototypes est qu'ils ne prennent pas en compte la variabilité du signal. Un geste donné n'est jamais exactement reproductible. Il faut tenir compte des variations du

signal qui peuvent avoir de multiples sources (utilisateur, capteur) et des variations inter-utilisateurs. Pour tenir compte de tous les cas possibles, il faudrait définir un prototype pour chaque combinaison de valeurs du vecteur de représentation, ce qui amènerait à faire n^m comparaisons, n étant le nombre de combinaisons et m la taille du vecteur.

Afin de diminuer la taille du vocabulaire, une approche dite statistique est parfois utilisée. L'approche statistique consiste à étudier la répartition des données en réalisant des calculs statistiques lors d'une courte phase d'apprentissage. Le calcul le plus simple est le calcul du *vecteur de paramètres de représentation moyen* sur plusieurs exemples d'une même posture [Da Silva Faria O. 1991]. Une autre méthode consiste à calculer, à partir des données, les regroupements de postures selon leur ressemblance. Ces regroupements sont utilisés pour définir des codes pour chaque posture [Takahashi T. et Kishino F. 1991], ou encore pour construire des arbres binaires de décision [Cagin J.-M. 1993].

Même avec une approche statistique, le vecteur de représentation obtenu n'est pas toujours très représentatif, car il est calculé sur un petit nombre d'exemples et la variabilité du signal n'est pas très bien représentée.

3.1.3.2. La discrétisation de l'espace en zones

La discrétisation de l'espace en zones est une méthode simple qui consiste à segmenter l'espace de représentation des données en zones délimitées par des valeurs fixées a priori. Cette division de l'espace en zones permet d'apporter une certaine souplesse dans la représentation du geste. Ils sont décrits comme faisant partie d'un intervalle de valeurs possibles, plutôt que par des valeurs uniques. Cela permet de diminuer la taille du vocabulaire.

Ces zones peuvent être utilisées pour représenter des postures ou des gestes. Dans le cas des postures, les zones peuvent correspondre à des intervalles, par l'intermédiaire de valeurs de flexion minimale et maximale pour chaque doigt [Zimmerman T. G., Lanier J. et al. 1987]. Elles peuvent correspondre aussi à un codage des flexions, tel que "*tendu*" et "*plié*" [Braffort A. 1992], [Revesz P. Z. et Raghava-Rao V. K. 1993], [Liang R.-H. et Ming O. 1995], complété parfois par un code "*indéterminé*" [Takahashi T. et Kishino F. 1991].

Dans le cas des gestes, les zones correspondent à des divisions de l'espace que parcourt la main sous forme de cubes [Searles D., Smith J. et al. 1993], ou à des divisions de l'espace que parcourent les doigts sous forme de postures clés [Bordegoni M. et Hemmje M. 1993].

Avec ce type de représentation, il est nécessaire de définir des valeurs limites, qui sont généralement fixées de manière arbitraire ou empirique. Si le geste subit une variation importante, il risque de se produire des erreurs au niveau de la représentation, qui vont fausser ensuite l'étape de décision. Ce type de représentation n'est pas très robuste à la variabilité du signal.

3.1.3.3. Les caractéristiques cinématiques et géométriques

Les caractéristiques cinématiques et géométriques sont des informations globales sur le geste nécessitant des calculs qui peuvent être plus ou moins complexes. Les caractéristiques cinématiques concernent la dynamique du geste (vitesse, accélération…), tandis que les caractéristiques géométriques concernent la forme du geste (rayon de courbure…).

Gestes 2D

Ce type de représentation est souvent utilisé en reconnaissance de geste 2D. Dans ce cas, les caractéristiques sont calculées sur la trajectoire représentant la trace du mouvement dans deux dimensions. Un des travaux de référence dans le domaine du geste de commande 2D est celui de Dean Rubine [Rubine D. 1991a], [Rubine D. 1991b]. Treize caractéristiques sont calculées, dont la longueur de la diagonale de la boite englobante, la distance entre le premier et le dernier point, la longueur totale de la trajectoire, l'angle traversé, la vitesse maximale, la durée.

La méthode de Rubine contient intrinsèquement une approche statistique. Chaque référence est constituée d'un vecteur de poids qui est calculé durant une phase d'apprentissage : pour chaque classe de gestes, la moyenne des caractéristiques est calculée, ainsi qu'une matrice de covariance. Lorsque l'apprentissage est fini, le vecteur de poids est calculé, pour chaque classe, en fonction des vecteurs de moyennes et des matrices. Les calculs détaillés sont donnés dans la thèse de Rubine au Chapitre 3 (p.50 à 52).

Gestes 3D

Dans le cas des gestes de la main, les caractéristiques cinématiques et géométriques peuvent être calculées sur les trajectoires de la main et des doigts dans l'espace.

Plusieurs travaux se sont inspirés des algorithmes de Rubine. Certains sont cités dans [Sturman D. J. et Zelter D. 1994]. Dans sa thèse, David Sturman a adapté les caractéristiques choisies par Dean Rubine pour les gestes 3D et les a modifiées de manière à pouvoir réaliser

une reconnaissance continue, sans points initiaux et finaux explicites [Sturman D. J. 1992]. Il a ajouté quelques caractéristiques choisies en fonction du type de gestes constituant le vocabulaire. L'interprétation des caractéristiques est réalisée en explicitant chaque geste et non pas à partir d'exemples fournis au programme d'apprentissage. Cette approche limite l'utilisation de l'outil de reconnaissance aux applications ayant le même vocabulaire gestuel.

Dans le cadre du stage de DEA [Braffort A. 1992], nous avons gardé les caractéristiques d'origine de D. Rubine auxquelles ont été ajoutées les caractéristiques suivantes : valeur maximale et distance parcourue sur le troisième axe, flexion maximale et angle parcouru pour chaque doigt (voir Chapitre 1). Ici encore, ces caractéristiques ne sont pas générales car elles sont liées au type d'interaction gestuelle utilisée dans l'application développée.

D'autres travaux ont utilisé le calcul de caractéristiques cinématiques et géométriques pour segmenter le signal. Celui-ci peut se réduire à des séquences de points caractéristiques tels que ceux pour lesquels la vitesse est minimale et ceux qui correspondent à un changement de direction [Sagawa H., Sakou H. et al. 1992], les points de rebroussement, les début et fin de segments. Il peut aussi se réduire à une séquence de segments de droites ou d'arcs [Bellalem N. 1995].

Dans [Kadous W. 1995], plusieurs essais de reconnaissance ont été effectués, sur des caractéristiques différentes, telles que la distance totale, l'énergie, le nombre de trames, la taille de la boite englobante, ainsi que le calcul d'histogrammes sur les caractéristiques précédentes. Mais cette étude ne fournit pas d'analyse ni de justification sur le choix des caractéristiques. Celles qui ont été gardées sont celles pour lesquelles le taux de reconnaissance est le meilleur (80% de taux de reconnaissance sur 95 gestes de l'AusLan, la langue des signes australienne).

Dans la plupart des travaux (mis à part le précédent), le vocabulaire est constitué de gestes artificiels, peu nombreux et assez spécifiques. Les caractéristiques utilisées sont adaptées au vocabulaire ou à l'application, ce qui ne permet pas leur utilisation pour des vocabulaires plus importants.

3.1.3.4. Conclusion

Les prototypes sont simples à mettre en oeuvre mais doivent être associés à une étape d'apprentissage portant sur suffisamment d'exemples si l'on veut bénéficier de connaissances

statistiques représentatives des données. Les prototypes sont très généraux et peuvent servir pour tout type de geste.

La discrétisation de l'espace en zones est simple à mettre en oeuvre, mais elle ne respecte pas le critère de continuité exposé au début de ce paragraphe. Ce type de représentation n'est pas suffisamment robuste par rapport à la variabilité du signal.

La représentation à base de caractéristiques cinématiques et géométriques est plus complexe à mettre en oeuvre car il faut tester puis choisir les caractéristiques les plus adaptées au signal à traiter. Dans toutes les études utilisant ce type de représentation, les caractéristiques sont choisies de manière empirique ou en fonction de la composition du vocabulaire. Elles ne sont a priori pas adaptées à un vocabulaire étendu comme par exemple celui qui constitue la langue des signes. Il faut disposer d'une bonne connaissance du "signal" gestuel pour pouvoir choisir les caractéristiques cinématiques et géométriques les mieux adaptées. Pour cela, l'étude réalisée sur la LSF (Chapitre 2), sera un élément de choix essentiel.

Une fois le signal brut acquis et représenté sous la forme d'un vecteur de paramètres de représentation, la seconde étape consiste à donner ce vecteur en entrée du module de décision. Les différents méthodes de décision sont présentés dans le paragraphe suivant.

3.1.4. LA DECISION

Le rôle du processus de décision est de déterminer la classe à laquelle appartient la forme captée en fonction du vecteur de paramètres de représentation. Parmi les différentes méthodes existantes, certaines ne fonctionnent qu'avec un type de représentation, d'autres sont plus générales. Les processus de décision les plus couramment utilisés en reconnaissance de gestes sont l'**approche linguistique**, la **comparaison de prototype** (template matching), les **réseaux connexionnistes**, la **comparaison dynamique** et les **modèles de Markov cachés**. Nous avons choisi de présenter les différentes techniques par ordre croissant de complexité du signal traité : postures, gestes isolés, gestes connectés, gestes enchaînés.

Certaines de ces méthodes sont généralement classées dans les approches dites structurelles (approche linguistique, comparaison avec un prototype, comparaison dynamique) [Miclet L. 1984], par opposition aux approches dites statistiques [Duda R. O. et Hart P. E. 1973]. Les réseaux connexionnistes sont généralement classés à part. Les modèles de Markov cachés comportent une double-approche structurelle et statistique. Un tiers des

travaux présentés ici sont basés sur la comparaison de prototypes et un autre tiers, sur les réseaux connexionnistes (voir Annexe 3).

3.1.4.1. Approche linguistique

Les approches linguistiques consistent à appliquer la théorie des automates et des langages formels au domaine de la reconnaissance de forme. Le processus de représentation fournit une séquence de lexèmes représentant les postures. Le processus de décision est constitué d'un ensemble de règles qui permet d'associer des suites de lexèmes.

Dans [Hand C., Sexton I. et al. 1994], les lexèmes sont constitués de huit configurations statiques, et les règles définissent six gestes qui sont des séquences de lexèmes, parfois associés à un mouvement rectiligne simple (vers la gauche, la droite, l'avant, l'arrière, le haut ou le bas). Aucune indication n'est donnée sur le passage des valeurs brutes issues du gant aux symboles utilisés dans les règles. Le taux de reconnaissance obtenu est médiocre (de 15% à 80% selon le geste).

Une telle approche a été testée aussi dans [Kadous W. 1995], simultanément avec une approche de type comparaison de prototype (présentée dans le prochain paragraphe) pour un même type de représentation à base de caractéristiques cinématiques et géométriques. L'approche linguistique a été abandonnée du fait de ces mauvaises performances (50% de taux de reconnaissance contre 80% pour la deuxième approche).

Cette méthode très rigide ne semble pas adaptée pour le traitement de données très variables telles que celles issues des gestes.

3.1.4.2. Comparaison de prototypes

La méthode par comparaison de prototypes est sans doute la méthode la plus simple. Elle est basé sur une fonction qui mesure les similarités entre l'entrée et les vecteurs de représentation de référence. L'entrée est classée comme étant un membre de la même classe que celle de la référence dont il est le plus semblable, ou le plus proche. En général un seuil de reconnaissance est défini, en dessous duquel l'entrée est rejetée car elle n'est pas assez proche d'une des références.

Si **n** est la taille du vecteur, $\mathbf{x_i}$ la $i^{ème}$ valeur du vecteur à classifier et $\mathbf{rèf_i}$ la $i^{ème}$ valeur du vecteur de référence, la fonction de similarité **f** peut être :

- la somme de la valeur absolue des différences. Le geste reconnu est celui pour lequel cette somme est minimum [Zimmerman T. G., Lanier J. et al. 1987], [Takahashi T. et Kishino F. 1991] :

$$f = \sum_{i=1}^{n} |ref_i - x_i|$$

- la différence x_i-ref_i, comparée à un seuil [Da Silva Faria O. 1991] :

$$f_i = ref_i - x_i$$
$$\forall i \in \{0,\ldots,n\}, f_i < \text{seuil}$$

- la somme des carrés des différences entre x_i et ref_i. Les vecteurs les plus semblables sont ceux pour lesquels la somme est minimale [Newby G. B. 1993] :

$$f = \sum_{i=1}^{n} (ref_i - x_i)^2$$

- la combinaison linéaire des paramètres de représentation :

$$f = \sum_{i=1}^{n} ref_i \bullet x_i$$

  Cela revient ici à calculer le produit scalaire entre deux vecteurs. Plus le produit est grand, plus les deux vecteurs sont semblables. Le geste reconnu correspond à la référence pour laquelle ***f*** est maximale et dépasse un certain seuil. Si la valeur maximale est inférieure à ce seuil, un symbole est renvoyé pour indiquer que le geste est inconnu. Dans [Rubine D. 1991a], les vecteurs de référence sont constitués de poids calculés en fonction des exemples fournis pendant l'apprentissage. Dans [Liang R.-H. et Ming O. 1995], les vecteurs de référence sont constitués de codages (0 pour tendu et 1 pour plié).

On notera que les taux de reconnaissance obtenus avec ce type de processus de décision varient de 65 à 80% et sont parmi les moins bons, comme souligné dans le tableau récapitulatif au Paragraphe 3.1.5.

Arbre binaire de décision

La méthode à base d'arbre binaire de décision peut être vue comme un cas particulier de comparaison de prototypes. Un arbre binaire de décision est un arbre dont les noeuds contiennent des opérations booléennes effectuées sur les données fournies par le processus de représentation. Les deux branches partant de ce noeud correspondent aux réponses *Vrai* ou *Faux*. L'évaluation des conditions sur les données est effectuée depuis la racine jusqu'à

rencontrer une feuille. Les feuilles contiennent les noms des références. La classification correspond à un chemin de l'arbre jusqu'à une feuille donnée. Les arbres binaires de décision peuvent être construits "à la main" [Braffort A. 1992] [Kadous W. 1995], ou être dérivés à partir d'exemples de l'entrée [Cagin J.-M. 1993]. Dans ces études, l'opération booléenne consiste à comparer une combinaison linéaire des valeurs données par le gant, et un seuil.

L'intérêt de cette approche est que l'espace de recherche diminue de moitié à chaque comparaison. Ainsi, la complexité est en $\Theta(\log(n))$ au lieu de $\Theta(n)$. Cependant, à chaque prise de décision, un risque d'erreur est possible. Donc le risque d'erreur est plus grand dans cette approche.

Les taux de reconnaissance obtenus peuvent toutefois être très bons pour des postures isolées [Cagin J.-M. 1993] (99%). Pour des gestes dynamiques, la seule étude disponible porte sur la saisie de gestes connectés à l'aide d'une caméra et obtient un taux médiocre de 45% [Tamura S. et Kawasaki S. 1988].

Conclusion

Les méthodes de type comparaison de prototypes sont faciles à développer et économes du point de vue informatique (temps de calcul, mémoire). Cependant, elles ne sont pas adaptées à la reconnaissance de gestes enchaînés. En effet, elles nécessitent une segmentation préalable du signal suivie de l'envoi de chaque geste segmenté au système de reconnaissance. La qualité de cette segmentation dépend du choix de seuils fixés a priori qui nuisent à la robustesse du système.

3.1.4.3. Réseaux connexionnistes

Les réseaux connexionnistes sont souvent utilisés en reconnaissance des formes. Un réseau connexionniste est composé d'un ensemble d'unités de calcul (neurones formels) reliées par des liens orientés (connexions) à travers lesquels circulent des valeurs numériques [Jodouin J.-F. 1994a]. Les liens sont caractérisés par une valeur numérique propre appelée poids synaptique. Chaque unité est un automate dont l'état (correspondant à la sortie) est donné par une valeur numérique appelée activation.

Les poids sont calculés durant une phase d'apprentissage. L'apprentissage est généralement supervisé : on utilise des patrons d'apprentissage (ou vecteurs de référence) comme activation des neurones de la couche d'entrée et on compare le résultat obtenu en sortie (activation des neurones de la couche de sortie) aux résultats attendus pour ce patron.

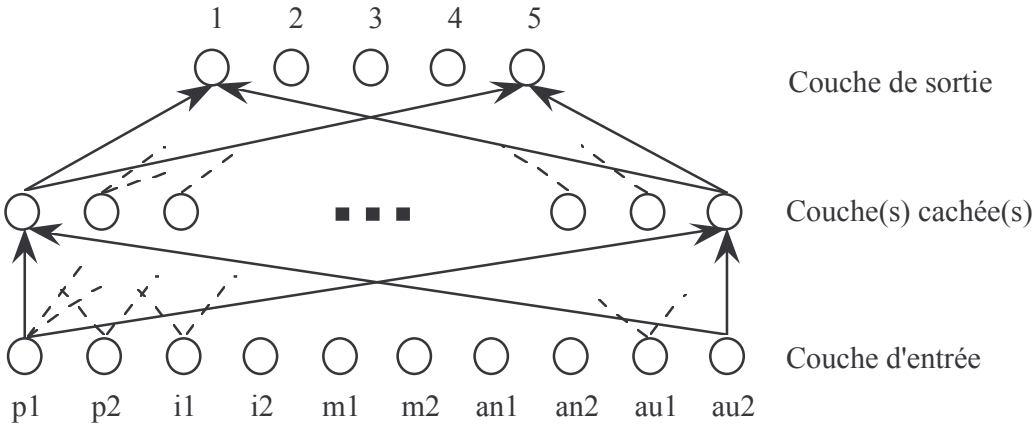
Cela permet de calculer une valeur d'erreur qui est utilisée pour mettre à jour les poids des neurones. Ce processus est répété jusqu'à ce qu'un état de stabilité du réseau soit obtenu. Suivant le cas, c'est le neurone de sortie dont l'activation est la plus grande qui désigne la forme reconnue, ou au contraire c'est l'ensemble des valeurs de la couche de sortie qui est significative.

Plusieurs types de représentation peuvent être utilisés, mais comme les valeurs données en entrée d'un réseau connexionniste doivent être normalisées (entre 0 et 1 ou -1 et +1), il y a toujours une phase de transformation des valeurs après passage dans le représenteur. La plupart des représentations utilisés sont à base de prototypes.

Selon que l'on cherche à reconnaître des formes statiques ou dynamiques, isolées ou enchaînées, les approches sont différentes. Parfois même, l'architecture du système est déterminée par l'application.

Reconnaissance de postures

La plupart des réseaux connexionnistes utilisés pour reconnaître des postures sont de type Perceptron Multicouches (PMC). Trois couches successives sont généralement utilisées, la couche d'entrée, la couche cachée et la couche de sortie. Chaque neurone d'une couche est relié à tous les neurones de la couche suivante (Figure 3.5).



p : pouce, i : index, m : majeur, an : annulaire, au : auriculaire

Figure 3.5 : Exemple de PMC sur un vocabulaire de 5 postures (chiffres de 1 à 5).

C'est à la couche d'entrée que sont fournies les données issues du processus de représentation, après normalisation. Elle possède un nombre de cellules égal à la taille du vecteur de représentation. Chaque unité est associée à une valeur du vecteur.

La couche cachée est de taille variable (choisie arbitrairement). En général des tests sont faits avec différentes tailles afin de déterminer la taille optimale à utiliser.

Le nombre d'unités de la couche de sortie correspond à la taille du vocabulaire dans [Kramer J. et Leifer L. 1989], [Vamplew P. 1993], [Murakami K. et Taguchi H. 1991], chaque unité est alors associée à un mot du vocabulaire. Dans [Gourley C. 1994], la couche de sortie comporte cinq unités, qui représentent cinq bits de codage permettant de représenter une posture. Afin d'obtenir une valeur binaire, un seuil (0,5) permet de décider si la valeur de l'unité est 0 ou 1.

Dans [Messing L. S., Erenshteyn R. et al. 1994], une approche originale est proposée. A partir d'une analyse linguistique, des groupes de ressemblance entre postures ont été définies. Une cascade de réseaux connexionnistes est utilisée (Figure 3.6). A chaque niveau de la cascade, des hypothèses sont éliminées. A la dernière étape, une unique solution est choisie. Le premier niveau permet de choisir un groupe parmi trois groupes de postures. Le deuxième niveau permet de choisir un sous-groupe parmi 3 sous-groupes de postures. Le dernier niveau permet de choisir une posture parmi trois (ou deux). L'intérêt de cette approche réside dans le fait que l'apprentissage est rapide du fait de la spécialisation des différents réseaux.

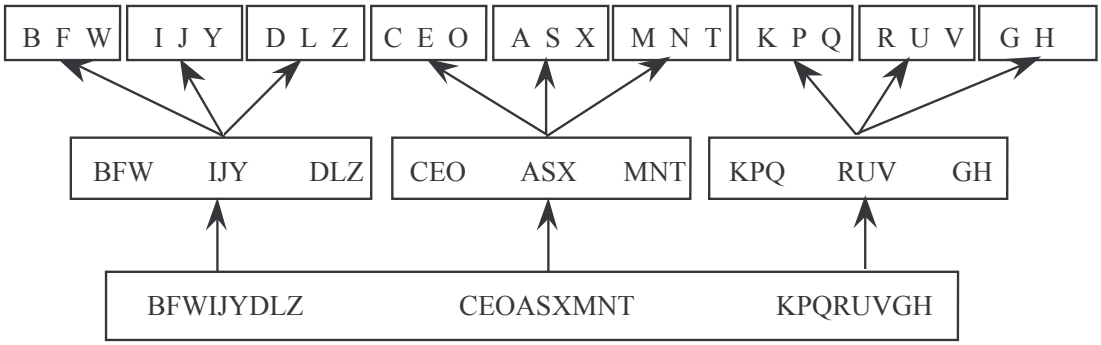


Figure 3.6 : Architecture de PMC en cascade.

Parmi toutes les études à base de PMC, les taux de reconnaissance obtenus varient entre 95 et 98% pour des corpus composés de postures isolées dont la taille varie de 5 à 42 postures. Les meilleurs résultats sont obtenus par [Murakami K. et Taguchi H. 1991] avec un

taux de reconnaissance de 98% sur un vocabulaire de 42 postures représentant l'alphabet manuel japonais, mais un temps d'apprentissage de plusieurs heures a été nécessaire sur une station Sun4.

Reconnaissance de gestes

Trois types de réseaux connexionnistes ont été utilisés dans le cadre de la reconnaissance de gestes dynamiques : les réseaux de type PMC, les réseaux de type Kohonen Features Maps (KFM) et les réseaux récurrents.

Réseaux de type PMC

Lorsque le signal à reconnaître est un geste, c'est-à-dire une succession de postures, la dimension temporelle intervient. Les PMC ne sont pas adaptés à ce type de données, c'est pourquoi les données doivent être normalisées sur l'axe du temps.

C'est le cas dans [Harling P. A. 1993] où les gestes sont normalisés sur un intervalle de **n** échantillons, afin d'obtenir une taille fixe pour la couche d'entrée du PMC. Dans ce cas, la taille de la couche d'entrée correspond à **n** fois la taille du vecteur de représentation. Le problème réside dans le choix du nombre **n** à cause de la variabilité des gestes dans le temps : un geste peut durer plus ou moins longtemps ou être exécuté plus ou moins vite.

Dans une autre étude effectuée au LIMSI [Collet C. 1993], les gestes sont enchaînés. Dans une première étape, ils sont segmentés et compressés en fonction de caractéristiques telles que les points de rebroussement, les zones de stabilité, et sont normalisés sur un intervalle temporel de sept vecteurs de paramètres répartis sur la durée du geste.

Durant les phases d'apprentissage et de reconnaissance, seuls les gestes formant le vocabulaire sont fournis au réseau après une suppression "manuelle" des gestes de coarticulation. En fixant un seuil d'activation, il serait possible d'éliminer automatiquement un certain nombre de gestes de coarticulation en sortie du réseau. C. Collet a constaté que plus ce seuil est élevé, plus on supprime de gestes de coarticulation, mais aussi plus on supprime de gestes du vocabulaire. En effet, les gestes de coarticulation sont parfois très semblables aux gestes du vocabulaire. C'est pourquoi il serait préférable de ne pas utiliser de seuil et d'envisager un traitement de plus haut niveau permettant d'interpréter la séquence complète de gestes.

Les résultats obtenus avec ce type de réseau sont de 91,4% pour la première étude [Harling P. A. 1993] portant sur des gestes isolés, et de 91,5% pour la seconde [Collet C. 1993] portant sur des gestes enchaînés mais après suppression "manuelle" des gestes de coarticulation.

Les défauts de cette approche sont d'une part la normalisation temporelle qui filtre les informations dynamiques portées par les gestes et d'autre part l'incapacité de traiter les gestes de coarticulation lorsque les gestes sont enchaînés.

Réseaux de type Kohonen Features Maps

Une autre approche consiste à utiliser des réseaux de type Kohonen Features Maps (KFM) [Boehm K., Broll W. et al. 1994]. Ces réseaux comportent deux couches correspondant chacune à une matrice (Figure 3.7). La première dimension de la couche d'entrée correspond à la taille du vecteur de données et la deuxième correspond au nombre de trames. Chaque unité de sortie correspond à un geste. La distance entre deux unités de la couche de sortie représente la mesure de similarité entre les deux gestes correspondants.

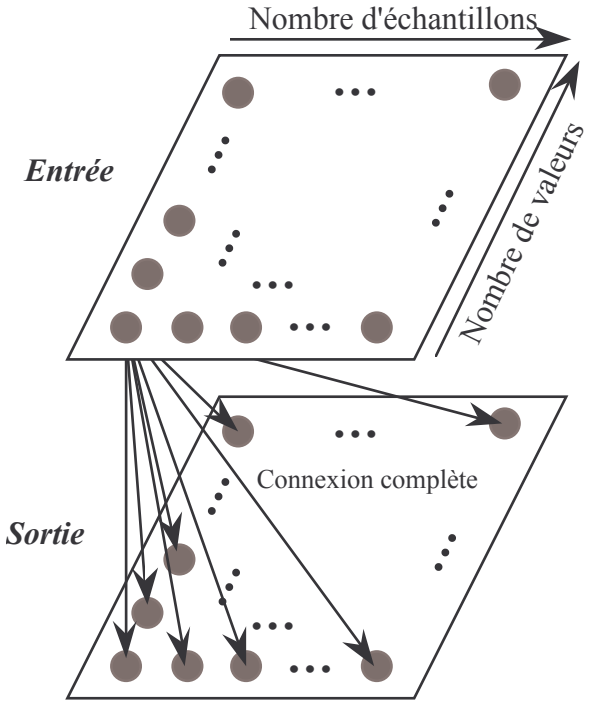


Figure 3.7 : Architecture des réseaux KFM.

De même que pour les réseaux de type PMC, le problème réside dans le choix du nombre de trames de la couche d'entrée. Pour résoudre cette difficulté, le système proposé

utilise plusieurs réseaux de type KFM travaillant sur des intervalles de temps différents. Malheureusement, aucun taux de reconnaissance n'est donné dans cette étude. De plus, de même que pour les PMC, ce type de réseau ne permet pas de traiter les gestes enchaînés.

Réseaux récurrents

Enfin, il existe un type de réseau adapté au traitement de données temporelles : les réseaux récurrents [Jodouin J.-F. 1994b]. Ils permettent de prendre en compte l'historique des données entrées, par l'intermédiaire d'une couche de contexte (Figure 3.8). Cette couche est une recopie de la couche cachée. Cela permet au réseau de connaître au temps t son propre état au temps t-1.

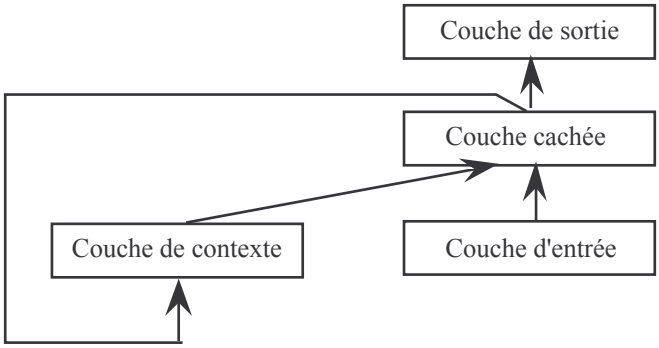


Figure 3.8 : Les différentes couches d'un réseau récurrent.

Ce type de réseau a été utilisé dans [Murakami K. et Taguchi H. 1991], associé à un réseau de type PMC chargé de détecter les postures initiales des gestes pour les segmenter.

Dans [Boehm K., Broll W. et al. 1994] un KFM et un réseau récurrent sont combinés. Les gestes sont segmentés à partir de caractéristiques telles que les points d'inflexion ou les points de rebroussement. Ces caractéristiques sont reconnues à l'aide du réseau de type KFM. Le deuxième réseau de type récurrent reconnaît le geste à partir d'une séquence de sorties du réseau KFM.

Le seul taux de reconnaissance fourni est celui de [Murakami K. et Taguchi H. 1991], qui est de 96% pour un vocabulaire de 10 gestes. Le temps d'apprentissage est de quatre jours sur une station Sun4.

Cette approche semble plus appropriée pour traiter des gestes dynamiques. Cependant, dans les deux études utilisant des réseaux récurrents, la segmentation entre les gestes enchaînés est traitée par un autre système. Il semble que ces réseaux ne permettent pas

facilement de traiter des gestes enchaînés et les gestes de coarticulation inhérents à ce type de corpus. De plus la durée de la phase d'apprentissage est en générale très importante, ce qui ne facilite pas la mise au point de ce type d'outil.

Approche de Fels et Hinton

Une des études les plus abouties est celle de Fels et Hinton [Fels S. S. et Hinton G. E. 1990], [Fels S. S. et Hinton G. E. 1993]. Le domaine d'application n'est pas la langue des signes. Il s'agit de développer un outil s'adressant aux personnes ayant perdu la possibilité de parler. Les gestes sont traduits en signaux qui sont ensuite synthétisés afin de produire de la parole. Le vocabulaire comprend des gestes qui constituent un langage artificiel. L'originalité de ce travail réside dans le fait que les potentialités du geste sont utilisées pour transmettre plusieurs informations en parallèle. Ainsi, la configuration de la main et la direction du mouvement servent à transmettre respectivement la racine et la terminaison des mots. De plus, la vitesse d'exécution du geste sert à déterminer la vitesse d'émission du mot et enfin, l'amplitude du geste permet de définir l'accentuation du mot. L'architecture utilisée est composée de cinq réseaux connexionnistes. Le premier est chargé de segmenter le signal. Les paramètres fournis au réseau sont la vitesse et l'accélération, ce qui permet de détecter les phases de décélération dans le mouvement. Les quatre autres sont dédiés à la configuration, la direction du mouvement, la vitesse et l'amplitude. Les performances du système sont bonnes (94%), mais le vocabulaire est trop limité pour pouvoir utiliser ce système efficacement (203 mots).

C'est pourquoi un deuxième système a été mis au point [Fels S. S. 1994]. S. Fels propose un système dans lequel les articulateurs gestuels sont mis en correspondance avec les articulateurs vocaux. Ainsi, les consonnes sont réalisées à l'aide de certaines configurations statiques, Les voyelles correspondent à d'autres configurations statiques associées à un déplacement dans le plan horizontal. Plus précisément, à chaque voyelle correspond une configuration de main ouverte et les valeurs X et Y de la position de la main dans l'espace correspondent à la position de la langue dans la bouche. Une main fermée correspond aux consonnes. Le petit doigt permet de spécifier si le son est voisé ou pas. Trois réseaux connexionnistes sont utilisés en parallèle. Le premier sert à différencier les voyelles des consonnes. Le deuxième sert à reconnaître les voyelles et le troisième est spécialisé dans les consonnes. Une pédale est utilisée pour la segmentation. Le système a été utilisé par une personne muette. L'outil semble long à prendre en main (100 heures d'entraînement) et le résultat sonore n'est pas très naturel (élocution lente et artificielle). Par contre, il est intelligible et l'utilisateur était satisfait de ne pas avoir de limitation du vocabulaire,

contrairement au premier système. Les performances globales ne sont pas données dans l'étude.

Conclusion

Du fait de l'étape d'apprentissage, les réseaux connexionnistes sont capables de reconnaître une forme bruitée ou incomplète et ils sont capables de généraliser. Les réseaux récurrents permettent de traiter des gestes enchaînés. On notera qu'ils ont été très majoritairement employés dans les différentes études existantes. Des outils permettant de les mettre en oeuvre sont très facilement disponibles [Erenshteyn R., Foulds R. et al. 1994].

Cependant, ils sont difficiles à mettre au point car le choix du nombre d'unités n'est pas connu a priori et de nombreux tests doivent être réalisés. De plus, si une nouvelle classe doit être ajoutée au vocabulaire, tout l'apprentissage est à refaire. Ceci peut être très long, surtout dans le cas où l'on manipule des réseaux récurrents.

3.1.4.4. Comparaison dynamique

Cette technique consiste à calculer une fonction de transformation temporelle permettant de faire coïncider deux signaux temporels. Si l'on considère un tableau 2D, et deux signaux à comparer (Figure 3.9), une fonction de transformation temporelle peut être vue comme étant un chemin dans le tableau.

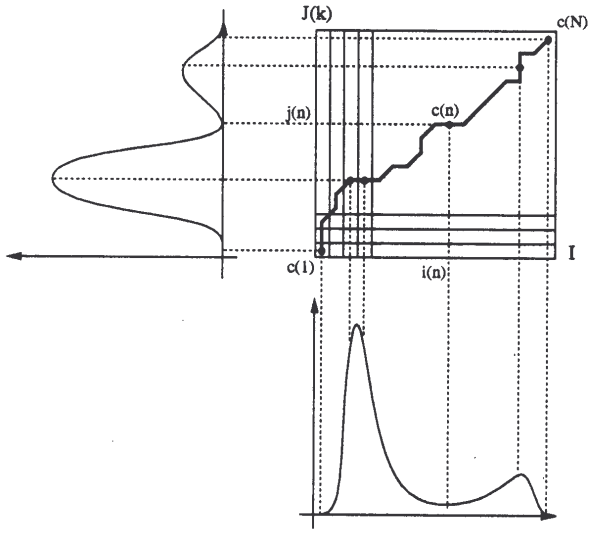


Figure 3.9 : Fonction de transformation temporelle.

Pour un signal de référence **k**, la fonction de coût $\mathbf{c_k}$ est définie par la séquence de couples $\mathbf{c_k(n) = (i_k(n), j_k(n))}$, pour **n** variant de **1** à **N**, **N** étant la longueur du chemin. Le coût du chemin est calculé à l'aide de distances euclidiennes. Le détail des calculs est donné dans [Cagin J.-M. 1993]. Le coût cumulé le long du chemin optimum représente l'indice de dissimilarité entre les deux signaux. Ainsi, pour classer un signal, on calcule le coût du chemin optimal entre ce signal et chaque signal de référence. Le signal de référence le plus proche de celui que l'on cherche à classer correspond au coût le plus petit.

Cette technique permet de traiter des gestes isolés dans [Da Silva Faria O. 1991]. Elle permet aussi de traiter des gestes connectés en utilisant une méthode récursive [Cagin J.-M. 1993]. La segmentation est réalisée lors de la reconnaissance et non pas au préalable comme c'est le cas des autres approches. Cela permet d'éviter l'utilisation de seuils arbitraires et la segmentation est plus robuste. Dans [Sagawa H., Sakou H. et al. 1992], les gestes sont enchaînés, mais aucune précision n'est donnée sur la manière dont la segmentation est calculée. Les taux de reconnaissance obtenus sont de 93,9% pour [Cagin J.-M. 1993], pour un vocabulaire de 14 mouvements connectés (construits à partir de 66 signes) et de 97,3% pour un vocabulaire de cinq phrases (composées à partir de 17 signes). Cette dernière valeur est meilleure mais porte sur un vocabulaire plus réduit.

Le défaut de cette technique est qu'elle est très gourmande en temps et en espace mémoire (surtout pour la version récursive), du fait du calcul du chemin optimal. Cependant, des optimisations algorithmiques peuvent être apportées [Cagin J.-M. 1993] (p.29 et 32-33), ou encore des compressions de données peuvent être appliquées [Sagawa H., Sakou H. et al. 1992].

La première technique ne permet pas de traiter des vocabulaires de gestes enchaînés. La deuxième le permet mais aucune indication n'est donnée sur la méthode utilisée. De plus, la compression des données implique le choix d'un seuil fixé a priori, ce qui semble être un frein à la robustesse du système. Ces deux techniques n'ont pas été retenues.

3.1.4.5. Modèles de Markov Cachés (HMM)

Les modèles de Markov cachés sont utilisés avec succès dans des domaines comme la reconnaissance de parole continue [Mariani J. 1993], [Gauvain J. L., Lamel L. et al. 1994], ou de prosodie [Geoffrois E. 1995]. Ils permettent de traiter des données dynamiques et enchaînées.

Une description formelle des HMM est donnée dans [Rabiner L. R. 1989]. Leur principe est exposé ici sur un exemple concret. Pour simplifier, on considère une simple posture, en l'occurrence celle qui correspond à la configuration nommé **C** (Figure 3.10).

Figure 3.10 : Posture **C**.

Les HMM permettent de modéliser une forme en contexte, c'est-à-dire en prenant en considération ce qui la précède et ce qui la succède. On suppose dans cet exemple que la posture **C** est précédée de la posture **5** et suivie de la posture **S** (Figure 3.11).

Figure 3.11 : Posture **C** en contexte (précédée de **5** et suivie de **S**).

Un modèle de Markov consiste en un nombre donné d'états. Ici, le modèle comporte trois états : le premier représente la coarticulation de **C** avec **5**, le second état représente la partie stable de la forme à modéliser **C** et le troisième état représente la coarticulation de **C** avec **S**.

A chaque état, est associée une probabilité de transition vers un autre état. Chaque état peut transiter sur lui-même. Cela permet de représenter la variabilité temporelle. Par exemple, la posture **C**, peut durer 1/10ème de seconde, correspondant à 5 transitions du deuxième état sur lui-même à une fréquence de 60 Hz. De la même manière, les transitions vers les configurations **5** et **S** peuvent être exécutées avec des durées variables. Ceci est représenté par les probabilités de transitions indiquées dans la Figure 3.12.

De plus, à chaque état est associé une fonction de distribution de probabilité (Figure 3.12). Cela permet de représenter la variabilité du signal, due à l'utilisateur (tremblement, hésitation…) ou au capteur (dérive…).

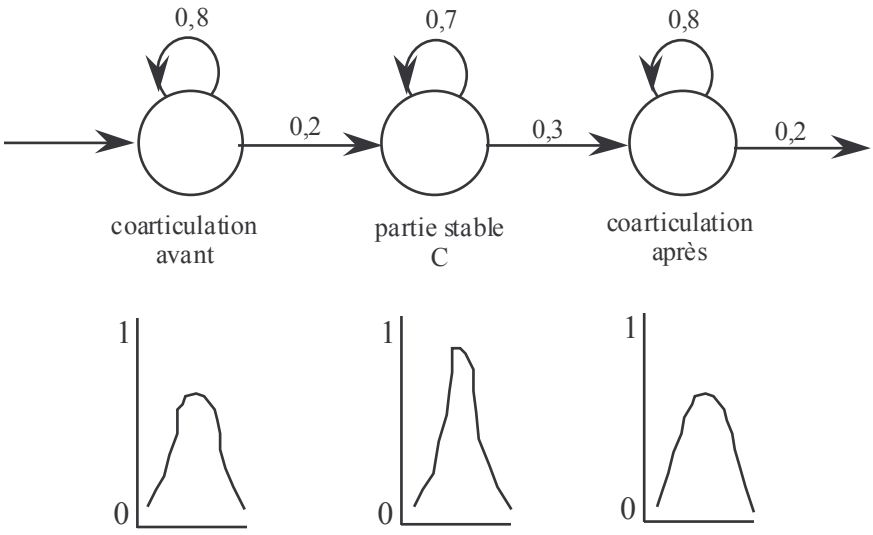


Figure 3.12 : Exemple de modèle de Markov caché pour la posture **C** en contexte.

Les différentes probabilités sont calculées durant une phase d'apprentissage. Les formules permettant l'apprentissage et la reconnaissance sont données en détail dans [Rabiner L. R. 1989] et [Starner T. E. 1995].

Un système de reconnaissance à base de HMM correspond à un graphe contenant tous les modèles désirés, chacun d'entre-eux représentant un élément du vocabulaire dans un contexte donné. Ces modèles sont connectés entre eux, ce qui permet de traiter des vocabulaires de gestes enchaînés.

Une seule étude, très récente, allie l'approche HMM et des capteurs de gestes de type gant numérique [Nam Y. et Wohn K. 1995]. Dans cette étude, un geste est défini comme étant composé de trois attributs parallèles : la configuration, l'orientation et le mouvement de la main. Une primitive de mouvement est définie comme étant un morceau de mouvement durant lequel aucune modification n'est observée dans la configuration et l'orientation. Un HMM est construit pour chaque primitive de mouvement et pour chaque coarticulation entre deux primitives. La segmentation entre les gestes enchaînés est cependant effectuée de manière ad hoc : elle est induite dans un premier temps par les variations de configuration et d'orientation lorsque ces variations dépassent des seuils fixés a priori, puis ensuite seulement, par les HMM. Cette approche est guidée par le type de vocabulaire étudié et de ce fait n'est

pas générale. Les taux de reconnaissance globaux ne sont pas donnés dans cette étude. Seuls les taux de reconnaissance pour chaque primitive de mouvement sont donnés (de 96,4 à 100%), hors segmentation.

Nous préférons l'approche utilisée par Thad Starner, dans laquelle aucun seuil n'est utilisé [Starner T. E. 1995], [Starner T. et Pentland A. 1995]. Le capteur utilisé est une caméra et le vocabulaire est constitué de gestes enchaînés des deux mains. De bons résultats sont obtenus (99,2%), avec l'aide d'une grammaire, sur des gestes pour lesquels la configuration n'est pas discriminante. En effet, comme cela a été indiqué au Chapitre 1, la capture de gestes par caméra ne permet pas d'obtenir suffisamment d'informations sur la posture des doigts.

Les modèles de Markov cachés permettent une double approche de reconnaissance, à la fois statistique et structurelle, par l'utilisation de règles de transition statistiques et en décomposant une forme donnée en éléments plus petits. Ils permettent de reconnaître des gestes enchaînés. Mais ils nécessitent la mise en place de gros corpus d'apprentissage de manière à pouvoir calculer correctement les fonctions de distribution de probabilité.

L'approche par HMM semble toutefois très prometteuse car elle est capable de prendre en compte l'ensemble des variabilités du signal sans imposer de traitement arbitraire préalable par seuillage ou équivalent.

3.1.4.6. Conclusion

Même s'il est difficile de comparer les performances des différentes techniques utilisées, du fait qu'aucun corpus commun n'existe, il est quand même possible d'avoir une idée de la robustesse relative des différentes techniques les unes par rapport aux autres. Comme cela a été indiqué au paragraphe précédent, dans les représentations de type caractéristiques cinématiques et géométriques, les caractéristiques qui dépendent de seuils doivent être évitées, afin que la propriété de stabilité soit préservée. De même, les processus de décision utilisant des seuils sont moins robustes que ceux qui n'en utilisent pas.

Si l'on travaille avec un vocabulaire de postures, la méthode par comparaison de prototypes et les réseaux de type Perceptron conviennent. Pour les gestes, les méthodes les plus prometteuses sont les réseaux de type récurrent, la comparaison dynamique et les HMM.

Les HMM permettent simultanément de segmenter et de reconnaître des phrases gestuelles sans imposer de traitement arbitraire préalable par seuillage, ce qui semble une

approche plus robuste. C'est une des raisons pour lesquelles nous avons choisi d'utiliser cette technique dans notre système de reconnaissance.

3.1.5. APPLICATIONS ET CORPUS

Ce paragraphe synthétise sous forme de tableaux les principales études présentées précédemment et pour lesquelles les performances sont indiquées. La première partie regroupe les études portant sur la reconnaissance de postures, tandis que la deuxième regroupe les travaux portant sur la reconnaissance de gestes.

Reconnaissance de postures

La reconnaissance de posture peut être basée sur un vocabulaire de configurations quelconques, mais la plupart du temps, elle est basée sur l'alphabet manuel.

L'alphabet manuel est composé de signes qui ont été créés afin de représenter les lettres de l'alphabet de la langue écrite du pays. Ainsi, en France, il existe 26 signes représentant les 26 lettres de l'alphabet. Ces signes diffèrent selon les pays. En France (LSF) et aux États-Unis (ASL), ils sont quasiment identiques. Ils sont réalisés à l'aide d'une seule main. Pour un seul signe, la configuration est dynamique (le **x**), pour un seul signe l'orientation est dynamique (le **j**) et pour deux autres signes, un mouvement est présent (le **y** et le **z**). Donc si l'on met de côté ces quatre signes, les 22 autres possèdent des paramètres complètement statiques. C'est pourquoi la grande majorité des études réalisées concernent l'alphabet manuel. Nous avons recensé une dizaine d'études. Certaines ne sont pas très précises, en particulier en ce qui concerne l'évaluation des outils de reconnaissance ou la composition du corpus. Seules les études pour lesquelles les informations sont complètes sont exposées ici.

Nom	Taille voc.	Type voc.	Capteur	Décision	Isolé/Ench.	Taux
Cagin	25	config. Stokoe[1]	DataGlove	ABD[4]	Enchaînés	99
Gourley	26	ASL[2]	CyberGlove	PMC[5]	Isolés	95
Harling	5	ASL[2]	PowerGlove	PMC[5]	Isolés	96
Messing	26	ASL[2]	CyberGlove	PMC[5] en cascade	Isolés	96,5
Murakami	42	JSL[3]	DataGlove	PMC[5]	Isolés	98
Takahashi	46	JSL[3]	DataGlove	Comp.[6]	Enchaînés	65

Tableau 3.1 : Reconnaissance de postures.

1 : configurations de base définies par Stokoe [Stokoe W. 1960]. Elles sont toutes différentes et statiques.
2 : American Sign Language
3 : Japanese Sign Language
4 : Arbre Binaire de Décision
5 : réseau connexionniste de type Perceptron multicouches
6 : comparaison de prototypes

On constate que toutes les études indiquées dans ce tableau obtiennent de bons résultats (de 95 à 98 % de taux de reconnaissance) lorsque les postures sont isolées. Notons que les systèmes de reconnaissance employés sont basés sur des réseaux connexionnistes de type PMC. L'approche par comparaison de prototypes est celle qui donne les moins bons résultats, sans doute parce que l'étude porte sur des lettres enchaînées.

L'étude fournissant les résultats les meilleurs est celle de Cagin, qui traite les configurations enchaînées et qui a utilisé un système à base d'arbre binaire de décision et une technique de segmentation par pause. Notons que dans cette étude, les erreurs dues à la segmentation ne sont pas spécifiées. Par ailleurs, le vocabulaire n'est pas constitué des lettres de l'alphabet, qui peuvent être ambiguës, mais de configurations statiques toutes distinctes, ce qui relativise les bons résultats obtenus.

Reconnaissance de signes

Dans le domaine de la reconnaissance de signes, les différentes langues actuellement étudiées sont l'ASL, la JSL, la LSF et l'AusLan (Australian Sign Language).

Nom	Taille voc.	Type voc.	Capteur	Décision	Isol/Conn/ Ench.	Taux
Cagin	14	prim. mouv.[1]	DataGlove	Comp Dyn[2]	Connectés	93,9
Cagin	66	ASL	DataGlove	ABD + Comp Dyn	Connectés	87
Collet	28	LSF	DataGlove	PMC	Enchaînés	90
Harling	3	ASL	PowerGlove	PMC	Isolés	91,4
Kadous	95	AusLan	PowerGlove	Caract./Comp.[3]	Isolés	80
Murakami	10	JSL	DataGlove	NN récurrent[4]	Enchaînés	96
Sagawa	17	JSL	DataGlove	Comp Dyn	Enchaînés	97,3
Starner	40	ASL	Caméra	HMM[5]	Enchaînés	97
Tamura	20	JSL	Caméra	ABD	Connectés	45

Tableau 3.2 : Reconnaissance de signes.

1 : primitives de mouvements
2 : comparaison dynamique
3 : comparaison de prototypes à partir de caractéristiques cinématiques et géométriques
4 : réseau connexionniste récurrent
5 : modèles de Markov cachés

Les trois meilleurs résultats sont obtenus par Murakami, Sagawa et Starner. Pour les deux dernières études, une grammaire a été ajoutée afin d'améliorer les performances du système de reconnaissance. Notons que le capteur utilisé par Starner est une caméra et de ce fait, la configuration des signes ne peut pas être distinguée précisément. Par contre, les deux mains sont captées, ce qui apporte des informations supplémentaires sur les signes et permet de travailler sur un vocabulaire plus étendu.

Reconnaissance de gestes quelconques

Beaucoup d'autres études ont porté sur des gestes artificiels, formant un vocabulaire de petite taille dédié à une application bien spécifique. Mise à part l'étude de Fels et Hinton présentée précédemment, ces applications portent en général sur l'utilisation du geste dans des applications multimodales, ou dans le domaine de la réalité virtuelle. Les méthodes de reconnaissance utilisées sont souvent ad hoc et ne peuvent pas être généralisables. De plus, les performances du système sont rarement données et les comparaisons entre ces différents travaux ne semblent pas possibles.

3.1.6. CONCLUSION

Une liste aussi complète que possible (une trentaine de références) des différents travaux réalisés dans le domaine de la reconnaissance de postures ou de gestes est présentée en Annexe 3, avec la spécification du capteur utilisé (gant ou caméra), des processus de représentation, de décision, de segmentation, la taille et le type du vocabulaire (posture/geste isolé/connecté/enchaîné), et enfin les performances, lorsqu'elles sont fournies.

Comme cela a déjà été indiqué, la robustesse des systèmes de reconnaissance est liée à la présence de seuils fixés a priori. Il est préférable d'éviter ces derniers.

Dans le cas où l'application utilise des gestes isolés, en nombre peu important, il est possible de se contenter d'un système tel que celui de Rubine, en prenant bien soin de choisir des caractéristiques cinématiques et géométriques qui ne nécessitent pas l'utilisation de seuils. L'apprentissage est beaucoup moins fastidieux que pour les réseaux connexionnistes et les HMM car le nombre d'exemples nécessaire par classe est faible (environ 15 d'après des essais faits par Rubine).

Si le type de vocabulaire envisagé comporte des postures ou des gestes enchaînés ou connectés, il est préférable de choisir un outil tel que les HMM, qui permet une segmentation en même temps qu'une reconnaissance, sans l'intervention d'aucun seuil. Cette technique permet de représenter les aspects de variabilités temporelle et spatiale, ainsi que le phénomène de coarticulation.

Pour les techniques basées sur la comparaison dynamique et les réseaux récurrents, la prise en compte du phénomène de coarticulation ne semble pas évidente. De plus, pour la comparaison dynamique, l'augmentation de la taille du vocabulaire accroît considérablement le temps de reconnaissance. Pour les réseaux connexionnistes la durée d'apprentissage est

plus longue que pour les modèles de Markov du fait qu'un grand nombre de cycles de calcul doit être effectué afin de diminuer la valeur d'erreur jusqu'à l'obtention d'une stabilisation. De plus, ils sont plus longs à mettre au point car la taille de la couche cachée n'est pas connue a priori.

Pour ces raisons, nous avons choisi de baser notre système de reconnaissance sur des modèles de Markov cachés. Reste le problème d'informations multiples évoqué au début de ce chapitre. L'architecture de notre système de reconnaissance et de compréhension devra tenir compte de cet aspect inhérent au canal gestuel.

Notons que dans le cas des gestes de langues des signes, l'utilisation d'une grammaire permet d'améliorer efficacement les performances du système [Starner T. E. 1995]. Toutefois, comme cela a été indiqué au Chapitre 2, l'ordre des signes est beaucoup moins significatif en LSF que l'arrangement spatial des gestes entre eux. De ce fait, les grammaires statistiques de type bigram ou trigram utilisées en reconnaissance de la parole continue ne sont pas adaptées à notre problématique. Ces grammaires permettent d'avoir une connaissance statistique sur les fréquences de succession de mots. En LSF, pour exprimer l'équivalent de "Le chat attrape la souris", les séquences [**chat**] [**souris**] [**attrape**] et [**chat**] [**attrape**] [**souris**] sont possibles, mais aussi [**souris**] [**chat**] [**attrape**] [Lane H., Boyes-Braem P. et al. 1976]. Même si certaines séquences sont plus souvent rencontrées que d'autres (dans le cas de notre exemple, l'ordre le plus courant est Agent-Patient-Action [Cuxac C. 1987]) il est probable que les grammaires statistiques ne suffiront pas. Cet aspect est développé plus en détail au Chapitre 4.

Comme nous nous intéressons en priorité aux phrases gestuelles de la LSF, nous avons constitué des corpus de signes enchaînés. Cela a nécessité de construire au préalable différents outils permettant d'une part de préparer les corpus d'apprentissage et d'autre part, de tester les paramètres de représentation adéquats. Ces outils sont présentés dans le chapitre suivant, ainsi que les limitations dues au système de capture utilisé.

3.2. OUTILS UTILISES ET DEVELOPPES

Ce paragraphe présente d'une part le gant numérique utilisé ainsi que ses limitations, d'autre part les différents programmes qu'il a fallu mettre en oeuvre dans le cadre de la reconnaissance de gestes.

3.2.1. LE SYSTEME DE CAPTURE DE GESTES ET SES LIMITATIONS

Comme cela a été indiqué au Paragraphe 1.1.3., les possibilités et les performances des systèmes de reconnaissance de gestes dépendent en grande partie du système de capture utilisé. En ce qui nous concerne, nous disposons d'un gant numérique DataGlove de VPL (Figure 3.13).

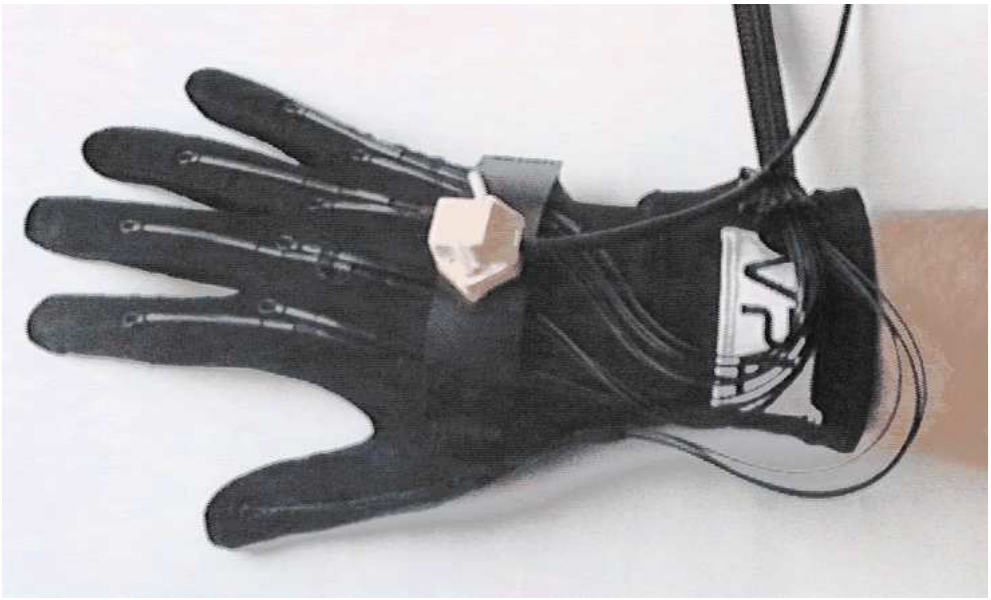


Figure 3.13 : Le DataGlove de VPL.

La première limitation à laquelle nous sommes confronté est que seul les gestes d'une main, la main droite, pourront être mesurés. Cependant, comme nous l'avons constaté durant l'analyse présentée au Chapitre 2, si l'on considère que la main dominante est la main droite, un seul gant permet de capter tous les signes à une seule main, soit plus d'un tiers des signes (voir Chapitre 2).

De plus, en ce qui concerne les signes à deux mains, un deuxième gant n'est sans doute pas obligatoire : connaissant le mouvement des deux mains, on peut en déduire les rapports entre les deux configurations (Chapitre 2). Il suffirait d'adjoindre un système de détection de mouvement (caméra ou Polhemus) pour obtenir les informations les plus importantes concernant la main gauche. Les signes effectués avec les deux mains pourraient alors être captés.

Ce gant est composé de deux systèmes de mesures indépendants. Un système à base de fibres optiques et un autre à base de bobines électromagnétiques.

3.2.1.1. Capture de la configuration

Le premier système permet de mesurer certaines valeurs angulaires de flexion des doigts à raison de deux valeurs par doigt. Les fibres optiques sont placées sur le dos de chaque doigt. Les propriétés optiques des fibres optiques permettent de différencier les postures du doigt par la conduction plus ou moins grande de la lumière en fonction de la courbure des fibres. Quand le doigt est tendu, le signal lumineux envoyé dans la fibre est peu atténué. Plus on fléchit le doigt, plus le signal lumineux traversant la fibre diminue. Pour associer des valeurs angulaires aux valeurs de flux lumineux, il est nécessaire de déterminer une fonction approximant la réponse des fibres.

Fonctions d'approximation

La réponse des fibres optiques à la flexion n'est pas linéaire. Il est possible d'approximer leur réponse par la formule suivante : $\mathbf{f(x) = a + b.x + c.ln(x)}$ [Roskos E. M. et Zhuang J. 1990]. Cela nécessite de déterminer au préalable les coefficients **a**, **b** et **c**, lors d'une phase dite de calibration.

A l'aide d'un système de mesure externe précis (à base de goniomètre), les valeurs de flexion ont été mesurées et comparées aux valeurs données par le gant avec la formule précédente. Les erreurs varient de 0 à 5 degrés, avec une moyenne de 1,8 degrés [Roskos E. M. et Zhuang J. 1990].

Malheureusement, la flexion répétée des fibres optiques provoque un vieillissement rapide et leur dynamique s'en trouve réduite. Il n'y a pas eu d'études portant sur le vieillissement des fibres optiques. Il semble à l'usage que ce soit les fibres les plus souvent sollicitées qui vieillissent le plus vite.

Nous avons développé plusieurs programmes de calibration. Le premier (nommé Cali1) approxime la réponse des fibres optiques avec une simple équation linéaire : $\mathbf{g(x) = ax + b}$. Durant la phase de calibration, le programme demande à l'utilisateur successivement de plier les articulations et de les étendre. Les valeurs extrêmes du flux lumineux ainsi obtenues sont mémorisées. Par la suite, les valeurs mesurées sont transformées en valeurs angulaires en calculant un simple rapport de proportionnalité par rapport aux valeurs extrêmes : $\mathbf{g(x) = (max-min)/90}$ pour une flexion variant de 0° a 90° (mis à part le pouce). Le deuxième programme (Cali2) utilise la fonction **f** présentée précédemment.

Aucune différence observable n'a été notée entre les performances des deux programmes, n'ayant pas de système de contrôle précis des valeurs angulaires. Il semble que même la fonction logarithmique soit insuffisamment représentative, car les fibres sont fixées sur le gant de manière assez rudimentaire et ne coulissent pas facilement, d'où des "à coups" dans les valeurs angulaires obtenues. A l'usage, il ne semble pas que l'on obtienne les 5° de précision annoncés dans [Roskos E. M. et Zhuang J. 1990], et sûrement pas le 1° annoncé dans le manuel de référence de VPL [VPL 1989]. Enfin, la qualité de la calibration dépend de la morphologie de la main utilisée. Il ne faut pas que la main soit trop petite, sinon le gant "glisse" dessus.

Durant les expérimentations, nous avons pu constater qu'une décalibration se produit assez rapidement, obligeant à recalibrer le gant régulièrement lors des séances de saisie de corpus. Si la calibration est réalisée avant la saisie, cette dérive provoque des saturations du signal, ce qui induit de gros problèmes lors de l'apprentissage car les valeurs obtenues sont artificiellement statiques. Il suffit alors, lors de la reconnaissance, que cette valeur diffère très légèrement pour que la reconnaissance échoue. Nous avons mis au point un processus de post-calibration, qui permet de prendre en compte cette dérive. Les valeurs maximale et minimale de chaque doigt sont calculées sur l'ensemble des mesures effectuées *avant*, *pendant* et *après* la saisie et non plus seulement *avant*. Toutes les erreurs provoquées par ce phénomène ont ainsi pu être éliminées.

Valeurs mesurées

Les symboles $\boldsymbol{\theta 1}$, $\boldsymbol{\theta 2}$ et $\boldsymbol{\theta 3}$ représentent des valeurs angulaires de flexion. Le symbole $\boldsymbol{\rho}$ représente l'abduction. Les symboles **x**, **y** et **z** représentent la position et les symboles $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ représentent l'orientation de la main.

Pour chaque doigt sauf le pouce, deux valeurs angulaires sont mesurées (Figure 3.14) : la flexion de l'articulation métacarpo-phalangienne **θ1** et la flexion de l'articulation inter-phalangienne **θ2**. L'angle entre la dernière phalange et la deuxième **θ3** n'est pas mesuré. Cela n'est de toute manière pas indispensable puisque des études ont montré que la valeur de cette articulation est entièrement corrélée aux valeurs des autres articulations. Une formule simple permet de calculer cette valeur. Si **θ3** est la valeur angulaire de flexion de l'articulation d'extrémité d'un doigt et **θ2** est celle de l'articulation inter-phalangienne, on a : **θ3 = 2/3 * θ2** [Lafouillade E. 1992].

En ce qui concerne le pouce, seuls **θ2** et **θ3** sont mesurés, ce qui n'est pas suffisant pour calculer **θ1** car les axes de flexion et de rotation du pouce ne sont pas triviaux, du fait que le pouce fait partie de la paume de la main et n'est pas parallèle aux autres doigts.

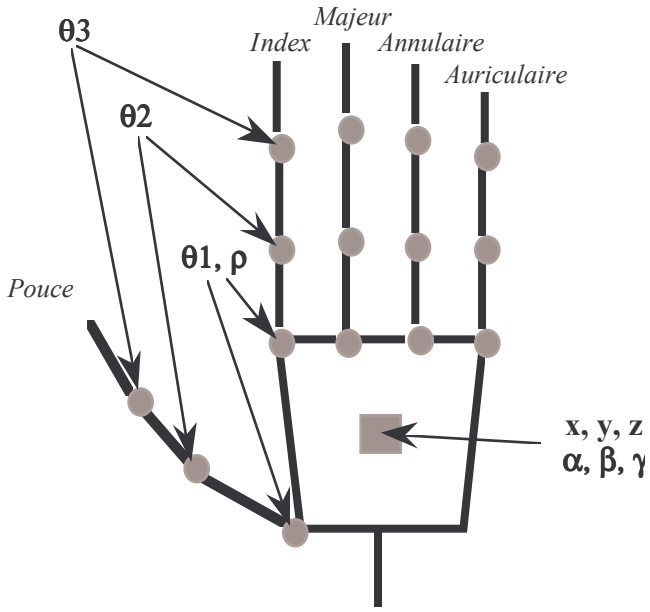


Figure 3.14 : Les degrés de liberté de la main.

Pour les programmes de calibration, les valeurs extrêmes fixées, pour chaque flexion mesurée par le gant, sont les suivantes :

flexion	pouce	flexion	index	majeur	annulaire	auriculaire
θ2	0/45	**θ1**	0/90	0/90	0/90	0/90
θ3	0/90	**θ2**	0/90	0/90	0/90	0/90

Tableau 3.3 : Valeurs extrêmes pour chaque flexion.

Ainsi, certaines informations importantes ne sont pas mesurées par ce gant :

- tout d'abord, la rotation du pouce (angles **θ1** et **ρ**), ce qui ne permet pas de différencier la position du pouce par rapport à la main. Par exemple, les configuration **5** et **5p** ne peuvent pas être différenciées. La configuration 5p est identique à la configuration 5, mis à part le pouce qui est placé devant la paume de la main (Figure 3.15) ;

Figure 3.15 : Les configurations **5** et **5p.**

- l'abduction (valeurs **ρ**). Par exemple, les configurations **5** et **moufle** (Figure 3.16) ne peuvent pas être différenciées.

Figure 3.16 : La configuration **moufle.**

- la flexion et la rotation du poignet. Nous ne disposons pas de système permettant de différencier une flexion du poignet du mouvement du bras provoquant la même trace dans l'espace. De même pour la rotation.

Ces limitations réduisent considérablement le nombre de gestes qu'il est possible de distinguer avec ce gant.

3.2.1.2. Capture de l'emplacement et de l'orientation

Le deuxième système de capture, le 3Space de Polhemus, complètement indépendant du premier, permet de capter l'emplacement et l'orientation de la main par rapport à un repère absolu fixe.

Il est composé de deux modules, chacun composé de trois solénoïdes disposés selon trois axes orthogonaux.

- Le premier module est un émetteur : un courant électrique parcourt les bobines et produit trois champs électromagnétiques orientés selon les trois axes. Ce module est placé à un endroit fixe et définit le repère d'origine.

- Le deuxième module est un récepteur : la réception des trois champs électromagnétiques produit trois courants induits par effet Hall dans les trois bobines. Les valeurs de ces trois courants induits sont proportionnelles à la position et l'orientation du module passif par rapport au module actif. Ce module est placé sur le gant. C'est lui qui est mobile.

Ce système permet d'obtenir directement les valeurs d'emplacement en centimètres et les valeurs d'orientation en radian ou en degré (Figure 3.14, valeurs **x**, **y**, **z** et α, β, γ). De plus, ce système est indépendant de la morphologie de l'utilisateur. Aucun programme de calibration n'est nécessaire.

3.2.1.3. Capture des informations dynamiques

Les informations dynamiques fournies par ces deux systèmes sont données par une succession de ces valeurs ponctuelles. Les fréquences d'échantillonnage des deux systèmes de mesure sont synchronisées et peuvent aller jusqu'à 60 Hertz.

Ainsi, le premier système de mesure peut fournir des valeurs de configurations statiques et dynamiques, tandis que le deuxième peut fournir des valeurs d'emplacement et d'orientation statiques et dynamiques. Remarquons que l'emplacement dynamique représente le paramètre de mouvement défini au Chapitre 2.

Ce qu'il est important de retenir à ce niveau est que la mesure du paramètre de configuration est complètement décorrélée de la mesure du paramètre d'emplacement (et de mouvement) et du paramètre d'orientation. Par contre les mesures des paramètres d'emplacement et de mouvement sont évidemment complètement corrélées puisque le mouvement est une succession d'emplacements. Enfin, le paramètre d'orientation est mesuré par rapport à un repère absolu et non par rapport à un système fixé sur l'avant-bras de l'utilisateur. De plus, on ne peut pas compléter l'information à l'aide d'un autre système qui donnerait la valeur angulaire de flexion et de rotation indépendamment du mouvement du bras. Pour ces deux raisons, le paramètre d'orientation n'est pas complètement décorrélé du paramètre d'emplacement.

De part la construction de ce gant numérique[1], seulement deux canaux de mesure indépendants et limités sont disponibles.

Enfin, précisons que nous ne disposons pas de systèmes permettant de capter la présence de contacts sur la main ou entre la main et une autre partie du corps.

3.2.2. VAG 2[2] : VISUALISATION ET ANALYSE DE GESTES

L'étude présentée dans le Chapitre 2 a porté sur un corpus "papier" puisque les gestes étaient issus d'un dictionnaire. Les gestes étaient représentés sous forme de dessin en deux dimensions. Des erreurs de perception étaient possibles dans le cas des gestes que nous ne connaissions pas au préalable.

Nous avons réalisé une application, nommée *VAG 2*, permettant en particulier de percevoir les gestes réels sous une autre forme. Les gestes sont captés au moyen du gant numérique. Ils sont stockés dans des fichiers et peuvent être étudiés ultérieurement. Une première fonction permet de revisualiser le geste saisi à l'aide d'une représentation fil de fer de la main. Une deuxième fonction permet de visualiser les courbes de variation des valeurs mesurées, c'est-à-dire les 10 valeurs angulaires des doigts, les coordonnées x, y z et les trois angles d'Euler. Elle permet aussi de visualiser les vitesses et les accélérations.

Quelques fonctions d'analyse ont été conçues, pour chaque comportement dynamique mesuré. Ainsi, pour le paramètre de configuration, une fonction permet de détecter les configurations dynamiques de type fermeture et les configurations dynamiques de type ouverture. Pour le paramètre de mouvement, une fonction permet de détecter les mouvements de type droite, arc ou cercle. Pour le paramètre d'orientation, une fonction permet de détecter les orientations dynamiques.

Bien sûr, la mise au point de ces fonctions nécessite le choix de seuils, ce qui pose des problèmes dus à la variabilité du signal. Mais cela permet d'obtenir quelques informations

---

1 Notons que le gant numérique CyberGlove (Virtual Technologies) possède des capteurs supplémentaires permettant de mesurer la rotation du pouce, les abductions et l'orientation de la main d'une manière indépendante du mouvement du bras. Un tel gant permet de capter beaucoup plus d'informations et d'augmenter la taille possible du vocabulaire.

2 Une première version, VAG 1, a été développée par Christophe Collet [Collet C. 1994].

non disponibles dans le corpus "papier", concernant par exemple les gestes de coarticulation présents entre chaque geste dans une séquence de gestes enchaînés.

Cet outil a aussi servi à mettre au point un filtrage des données. Le filtre utilise une fonction de Poisson. Ce type de filtre est mieux adapté qu'un filtre de type gaussien pour le signal gestuel provenant du gant. En effet, il évite de trop délocaliser les points traités (les voisins immédiats ont moins d'influence). Les petites variations à haute fréquence dues au tremblement de la main ou au bruit du Polhemus sont filtrées. La mise au point est faite "à la main", par superposition et comparaison des courbes filtrées et non filtrées. La formule de la fonction de Poisson est : $f(x) = \exp(-|x - \tau| / B)$, avec τ représentant la moitié de la fenêtre et B représentant la largeur à mi-hauteur de la fonction. Les valeurs suivantes : $\tau = 4$ et $B = 5$ ont été utilisées. Ces valeurs correspondent approximativement au nombre d'échantillons qui sont nécessaires pour représenter une transition du signal.

Enfin, une autre fonction a été ajoutée afin de stocker dans le fichier de gestes les début et fin de gestes, ainsi que les labels du signe et de chaque paramètre. Cette segmentation est réalisée manuellement, en fonction de l'image fil de fer et des courbes de valeur. Les fichiers de gestes segmentés peuvent ensuite être utilisés dans la phase d'apprentissage du système de reconnaissance.

Notons que cet outil a nécessité la programmation d'interfaces X et Motif assez complexes. La Figure 3.17 montre l'outil VAG2. Les courbes sont en réalité en couleur afin de pouvoir les distinguer facilement. Un type de couleur est associé à chaque doigt et l'intensité (clair, moyen, foncé) indique si la courbe concerne l'articulation inter-phalangienne, métacarpo-phalangienne ou la somme des deux. L'interface de VAG2 est écrite en Motif/X11. Il fonctionne sous Unix sur station HP9000/715.

Figure 3.17 : VAG2.

3.2.3. TEPA : OUTIL DE TEST DES PARAMETRES

Comme cela a été indiqué au Paragraphe 3.1, le choix des paramètres de représentation est une importante étape lors de la conception d'un système de reconnaissance. Afin de pouvoir choisir les paramètres les plus discriminants, un outil nommé TePa a été développé. Il est composé de deux modules. Le premier est dédié à la configuration. Les paramètres recherchés sont fonction des valeurs angulaires des doigts. Le deuxième module est dédié au mouvement et les paramètres sont ici fonction des valeurs de position x, y, z.

Chacun des modules est en fait un système de reconnaissance de gestes isolés. La représentations est de type *caractéristique cinématiques et géométrique* et la décision est de type *comparaison de prototypes*, comme dans le système de Rubine [Rubine D. 1991a].

3.2.4. AUTRES OUTILS

Plusieurs autres outils ont été développés, tels que les programmes de calibration cités précédemment (Cali1 et Cali2), des programmes de gestion du port série sur lequel est branché le gant numérique qui permettent de fixer la fréquence d'échantillonnage, un programme de saisie de corpus de gestes (nommé SaiCoS) et enfin, un programme permettant de transformer les données sous une forme utilisable par l'outil de reconnaissance utilisé (nommé PrepaHMM)[3].

[3] L'ensemble de ces programmes représentent plus de 14000 lignes de code, écrits en C, C++ et Motif/X11.

3.3. EXPERIMENTATIONS REALISEES

Ce paragraphe décrit les expérimentations réalisées en reconnaissance de gestes enchaînés. L'architecture du système est très générale, c'est-à-dire qu'elle est indépendante du capteur et du système de reconnaissance utilisés. Elle tient compte de la structure et du rôle des paramètres qui forment les signes en LSF. Ce premier prototype permet de traiter les types de signes suivants : signes standard, classificateurs, verbes directionnels [Braffort A. 1996a], [Braffort A. 1996b].

Après avoir précisé les objectifs fixés, c'est-à-dire la reconnaissance de phrases gestuelles de la LSF, composées de signes de types différents, la composition du système de reconnaissance est étudiée. Il faut voir le système proposé comme un premier prototype, car il est pour l'instant limité à des phrases composé de signes n'utilisant qu'une main, ce qui réduit le type de signes que l'on peut traiter. Le corpus mis en place pour évaluer ce prototype est détaillé dans la deuxième section. Il sera aussi utilisé pour évaluer le prototype de système de compréhension présenté au Chapitre 4. Le système de reconnaissance est basé sur la mise en oeuvre de HMM. Le choix du nombre d'états des différents HMM dépend de la structure interne des signes constituant le vocabulaire. C'est donc après la présentation du corpus que sont détaillés les HMM représentant le vocabulaire choisi. Finalement, les premiers résultats d'évaluation sont présentés.

3.3.1. LE MODULE DE RECONNAISSANCE

Les corpus que l'on souhaite traiter sont constitués de phrases de la LSF, c'est-à-dire de gestes enchaînés. Ces phrase peuvent être constituées de signes standard, de classificateurs et de verbes directionnels. La Figure 3.18 montre un exemple de phrase gestuelle signifiant "*Je donne un gâteau au garçon qui est à ma droite*". Les signes [**garçon**] et [**gâteau**] sont standard, le signe "index vertical" est un classificateur et le signe [**donner**] est un verbe directionnel.

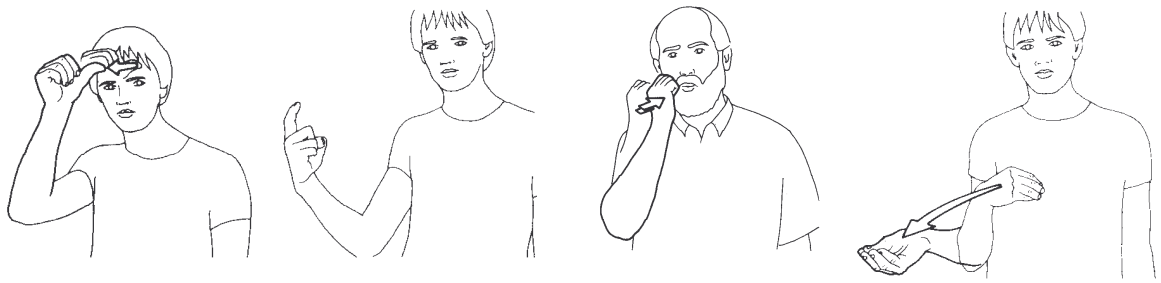


Figure 3.18 : Exemple de phrase en LSF composée des signes
[**garçon**], "index vertical", [**gâteau**] et [**donner**].

Comme nous l'avons exposé dans le Chapitre 2, deux catégories de signes peuvent être distinguées : ceux pour lesquels les quatre paramètres sont invariables quelque soit le contexte et ceux pour lesquels au moins un des paramètres est variable en fonction du contexte. La première catégorie correspond aux signes standard. La seconde inclut les classificateurs et les verbes directionnels.

Notre objectif est de pouvoir reconnaître ces différentes catégories de signes. Notons que dans toutes les études rencontrées (voir Paragraphe 3.1), seuls les signes standard sont traités. Ces signes sont plus simples à traiter car tous leurs paramètres sont invariables et il n'en existe qu'une seule "instance". Mais il est rare que dans une situation réelle, deux personnes sourdes n'emploient que des signes standard pour dialoguer. Comme cela a été indiqué dans le Chapitre 2, des signes non standard tels que les classificateurs sont souvent employés à titre de pronoms et un certain nombre de verbes usuels sont directionnels.

3.3.2.1. Composition du module

En fonction du comportement de leurs paramètres, trois catégories de signes sont distinguées, les signes standard, les signes standard variables et les signes non standard. Leurs principales caractéristiques sont rappelées ici.

- *Signes standard*

  Il s'agit des signes "prédéfinis" qui peuvent être répertoriés dans les dictionnaires. Ces signes sont utilisés pour représenter des entités, des adjectifs, des verbes non directionnels…

  Leurs quatre paramètres (configuration, mouvement, orientation et emplacement) sont définis de manière unique. Il n'en existe qu'une seule "instance".

- *Signes standard variables*

  Il s'agit des signes standard dont au moins un des paramètres peut varier en fonction du contexte. C'est le cas par exemple des verbes directionnels.

  Les verbes directionnels se conjuguent dans l'espace. La direction du mouvement et l'orientation de la main permettent de déterminer les rôles d'agent et de patient. Seuls la configuration et une partie du mouvement, la primitive de la trajectoire du mouvement définie au Chapitre 2 (statique, droite, arc, cercle), sont indépendants de la conjugaison. Parfois, le verbe peut intégrer un classificateur qui jouera le rôle de super-pronom (voit Chapitre 2) et dans ce cas, seule la primitive de mouvement est invariable.

- *Signes non standard*

  Ce sont des signes qui sont créés durant le discours, en fonction des besoins et du contexte. Un exemple de signe non standard est le classificateur.

  Un classificateur est un signe qui décrit et représente toute une classe d'objets par l'intermédiaire de leur forme. Ils ont une fonction de "super-pronom", c'est-à-dire qu'ils sont utilisés pour décrire la forme des entités et pour les localiser dans l'espace. Les paramètres de mouvement, d'orientation et d'emplacement varient selon le contexte. Seule la configuration est indépendante du contexte.

Pour les signes standard, les quatre paramètres peuvent être donnés en entrée du système de reconnaissance puisqu'ils possèdent chacun une unique valeur (modulée en fonction de la variabilité du signal due à l'utilisateur ou au capteur).

Par contre, les signes standard variables et non standard ne peuvent pas être traités de la même manière, car pour au moins un de leur paramètre, la variabilité du signal va être très importante, puisque sa valeur dépend du contexte. Par simplification, ce type de signe sera appelé *signe variable* dans la suite.

En conclusion, si l'on considère les verbes directionnels et les classificateurs, les seuls paramètres invariables quelque soit le contexte sont respectivement la primitive de mouvement et la configuration.

Une première version de l'architecture du système de reconnaissance a été étudiée et testée. Suite aux problèmes rencontrés, une deuxième version répondant aux principaux défauts précédemment rencontrés a été mise en place.

Première version

La première proposition est un système de reconnaissance comportant trois sous-modules fonctionnant en parallèle :

- le premier dédié aux signes standard,
- le deuxième dédié aux configurations,
- le troisième aux primitives de mouvement.

Nous avons procédé à une série d'essais à partir d'un corpus simple sur les modules concernant les configurations et les primitives de mouvement, et nous avons constaté quelques erreurs de reconnaissance pour le premier et beaucoup pour le second. Certaines de ces erreurs sont facilement explicables. En effet, le fait de décomposer un geste en parties qui vont être traitées de manière indépendante peut provoquer différents types d'erreurs :

- Si deux signes ayant la même configuration se succèdent dans une phrase gestuelle, il est possible qu'une erreur de type suppression se produise : le système de reconnaissance ne va pas pouvoir segmenter les deux signes. Le même problème peut apparaître avec le module dédié aux primitives de mouvement.

  C'est d'ailleurs pour ces mêmes raisons qu'il a semblé préférable de faire une reconnaissance globale des signes standard où tous les paramètres sont pris en compte simultanément, plutôt qu'une reconnaissance par paramètres.
- D'autre part, la primitive de mouvement n'est pas une information suffisamment discriminante pour plusieurs raisons :
  - Une de ses valeurs possibles est "immobile" (dans cette expérimentation, les valeurs considérées sont "mobile" et "immobile"). Or, parfois, le signeur ralentit son mouvement entre deux signes jusqu'à atteindre une vitesse nulle, afin d'effectuer une "petite pause" (inconsciente) entre deux signes. De ce fait, beaucoup d'insertions de la valeur "immobile" sont engendrées.
  - Certains signes possèdent un petit mouvement dont la vitesse est faible. D'autre part, la transition entre deux signes peut engendrer des mouvements dont la vitesse est élevée. Les vitesses correspondant aux valeurs "mobile" et "immobile" sont très variables et de ce fait, certains signes dont la vitesse est faible sont reconnus comme étant immobiles.
  - Nous avons constaté que les coarticulations entre les configurations et les primitives de mouvement n'étaient généralement pas synchronisées. Souvent, la transition au sein de la configuration se produit avant celle apparaissant dans le

mouvement. Il parait difficile ensuite de synchroniser les sorties des deux modules, même si la reconnaissance a donné des résultats corrects.

Il faudrait étudier ce phénomène plus en détail pour voir s'il s'agit d'un décalage systématique ou si cela dépend du signeur. Il est peut-être dû au fait que les doigts ont moins de distance à parcourir, même dans le cas d'une ouverture ou d'une fermeture complète des doigts.

Notons que le problème de synchronisation entre les différents articulateurs mis en jeu est crucial aussi dans le domaine de la génération automatique de gestes naturels. L'outil VAG2 présenté au chapitre précédent peut permettre aux chercheurs travaillant dans ce domaine de faire des analyses à partir de gestes réels.

Deuxième version

Pour ces raisons, nous proposons dans une deuxième version de regrouper en un seul module les données concernant la configuration et la primitive de mouvement. Ce module est ainsi dédié aux signes variables. Ainsi, le système de reconnaissance comporte deux sous-modules fonctionnant en parallèle :

- le premier dédié aux signes standard,
- le deuxième dédié aux signes variables.

Par rapport à la version précédente, la différence est que la taille **t** du vocabulaire d'apprentissage de ce module sera plus grand que la somme des valeurs **n** et **m** représentant respectivement la taille du vocabulaire de configurations et celle du vocabulaire de primitives (au pire, **n*m**). Cependant, cette valeur **t** restera bien plus petite que celle correspondant à la taille du vocabulaire de signes standard, donc cet aspect ne représente pas véritablement un problème.

Le regroupement des modules *configuration* et *primitive de mouvement* ne permet pas de résoudre tous les problèmes inhérents à cette architecture. Les différents aspects qui vont influencer les performances du système sont indiqués ci-dessous.

Liens entre les deux modules

Comme le montre l'exemple donné au début du Paragraphe 3.3, une phrase en LSF peut être constituée de signes standard, de classificateurs et de verbes directionnels. Le module dédié aux signes variables est capable de fournir une classification pour tout type de geste,

standard ou pas, puisque tout geste est constitué en particulier d'une configuration et d'une primitive de mouvement. Par contre, le premier sous-module dédié aux signes standard ne pourra pas reconnaître les signes qu'il n'a pas appris, ou du moins, il proposera une solution approchante mais incorrecte. Par conséquent, il doit pouvoir indiquer que le signe est "inconnu" ou "douteux" s'il ne reconnaît pas un signe standard.

Il faut pouvoir disposer, en sortie de ces modules, des scores de reconnaissance pour chaque élément de la phrase gestuelle. Ainsi, pour chaque signe, en comparant les deux scores, la sortie d'un des deux modules peut être choisie. A partir des deux séquences de symboles S1 et S2, on doit obtenir par l'intermédiaire de ce processus de choix une unique séquence de symboles S (Figure 3.19).

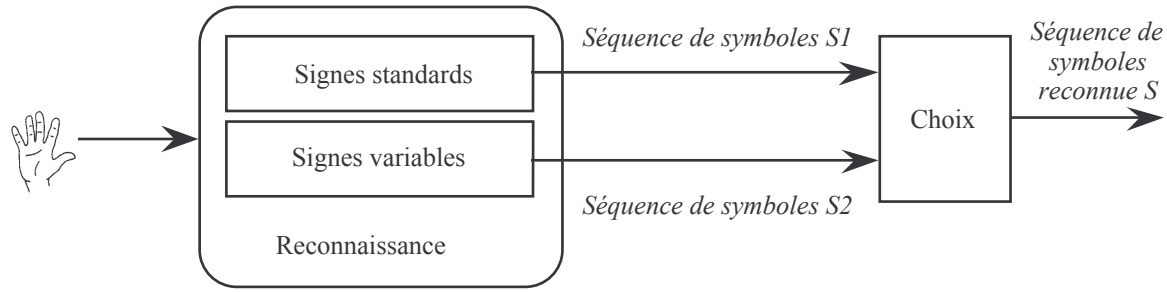


Figure 3.19 : Les deux modules du système de reconnaissance.

Il n'est pas évident a priori que les deux modules fournissent des séquences segmentées de manière identique. Une analyse des résultats obtenus sera nécessaire, afin de déterminer plus précisément le fonctionnement du processus de choix entre les deux sorties.

Ces deux modules sont composés, comme tout système de reconnaissance, de processus de *représentation* et de *décision*.

3.3.2.2. Paramètres de représentation

Aussi bien dans le premier que dans le second module, plusieurs informations ayant la même importance sont portées par le geste. Le premier module gère les informations portées par les quatre paramètres "linguistiques" du signe (configuration, mouvement, orientation et emplacement), réduits à deux paramètres "articulatoires" qui sont les données représentant la configuration (10 valeurs fournies par le DataGlove) et les données représentant l'emplacement et l'orientation (3 + 3 valeurs fournies par le Polhemus).

Une représentation à base de prototypes a tendance à donner beaucoup plus d'importance à la configuration qu'aux autres paramètres. Nous avons observé ce phénomène en analysant les causes d'erreurs du système de reconnaissance développé par C. Collet [Collet C. 1993]. Le corpus est composé de groupes de signes formant des paires minimales, c'est-à-dire dont un seul paramètre permet de les distinguer. La plupart des erreurs de reconnaissance, de type substitution, portent sur des signes qui ont une configuration identique.

Dans une des études citées précédemment [Murakami K. et Taguchi H. 1991], aux valeurs fournies par le gant, (posture, orientation et position absolue), a été ajoutée la position relative par rapport au début du geste. De plus, les positions absolue et relative sont fournies sur trois échelles différentes. Ainsi, il y a dix valeurs pour la configuration et dix-huit valeurs pour la position. Les performances du système (à base de réseau récurrent) sont bonnes (96 % de taux de reconnaissance) mais le corpus est constitué de seulement dix gestes segmentés au préalable et l'apprentissage dure quatre jours sur une SUN4.

Il faut choisir avec précaution à la fois le type de paramètre de représentation qui sera fourni au processus de décision, mais aussi leur répartition sur les paramètres du signe.

L'étude réalisée et présentée au Chapitre 2 a permis de connaître le comportement des paramètres de la LSF liés à la main et d'en déduire les paramètres de représentation les mieux adaptés.

Représentation des signes standard

Comme cela a été exposé précédemment, deux sources d'informations décorrélées sont fournies par le gant : la configuration et l'ensemble position/orientation.

La configuration peut être statique ou dynamique et dans ce cas, l'annulaire et l'auriculaire ont un rôle passif (voir Chapitre 2). Pour distinguer les configurations statiques entre elles, il faut disposer des valeurs angulaires fournies par le gant après calibration. Pour distinguer les configurations dynamiques simples, ouverture et fermeture, il faut considérer la vitesse angulaire pour chaque doigt sauf l'annulaire et l'auriculaire. Cette vitesse sera positive en cas de fermeture et négative en cas d'ouverture.

Pour l'ensemble position/orientation, le gant fournit 6 valeurs, trois pour la position et trois pour l'orientation. Ces valeurs sont absolues, par rapport à un repère fixe placé devant le signeur. Pour représenter les variations durant un signe, il faut fournir au système de

reconnaissance les valeurs, mais aussi les vitesses. Notons que les valeurs de l'orientation ne sont pas continues puisqu'elles varient circulairement de 0° à 360°, pour repasser brutalement à 0°. Afin de conserver la propriété de continuité que doit respecter un paramètre de représentation, il faut passer par des fonctions continues représentant ces valeurs angulaires, comme le sinus et le cosinus des angles.

Pour les signes standard, les paramètres de représentation sont :

- 10 valeurs de flexion des articulations des doigts,
- 6 vitesses de flexion des articulations des doigts,
- 3 valeurs de position,
- 3 vitesses par rapport à la position.
- 6 valeurs d'orientation,
- 6 vitesses par rapport à l'orientation.

Rappelons que les deux sources d'information doivent avoir le même poids. Ici, il y a 16 valeurs pour la configuration, 6 valeurs pour la position et 12 pour l'orientation, ce qui risque de poser des problèmes si des signes ne diffèrent que par le mouvement.

Représentation des signes variables

Ce module utilise la configuration et la primitive de mouvement.

Pour la configuration, les paramètres sont les mêmes que pour le module précédent.

Les primitives de mouvement possibles sont "statique", "droite", "arc", "cercle" et, plus rarement, "vague", "zigzag" ou autre (voir Chapitre 2). La norme de la vitesse est utilisée pour distinguer les signes "mobiles" des signes "immobiles". La courbure est utilisée pour distinguer les différentes primitives.

Pour ce module, les paramètres de représentation sont :

- 10 valeurs de flexion des articulations des doigts,
- 6 vitesses de flexion des articulations des doigts,
- 1 valeur représentant la norme de la vitesse.
- 1 valeur représentant la courbure.

Ici aussi, les deux sources d'informations doivent avoir le même poids. Or il y a 16 valeurs pour la configuration et 2 valeurs pour la primitive de mouvement, ce qui risque de poser des problèmes de déséquilibre encore plus grands que pour le module précédent.

Pour les deux modules, l'analyse des erreurs de reconnaissance indiquera si le système est sensible à ce déséquilibre ou pas. Cette analyse est présentée à la fin du chapitre.

3.3.2.3. Les processus de décision

Pour construire le système de reconnaissance, nous avons utilisé les programmes développés au LIMSI par Jean-Luc Gauvain pour la reconnaissance de parole continue. Ce système, à base de modèles de Markov cachés, est tout à fait représentatif de l'état de l'art dans le domaine de la parole et donne de bons résultats. Les modèles sont des chaînes de Markov gauche-droite, avec des densités de probabilité de type mixture de gaussiennes [Gauvain J. L., Lamel L. et al. 1994]. Les principes utilisés dans le mécanisme de reconnaissance de ces programmes sont très généraux. Ces derniers ont pu être appliqués aux gestes sans autre modification que les fichiers de données[4].

Comme cela a été expliqué à la Section 3.1.4.5., le nombre d'états des HMM représente la structure interne des signes. La description détaillée des différents HMM mis en oeuvre est reportée au Paragraphe 3.3.3, qui suit la description du vocabulaire constituant le corpus utilisé.

Par rapport à nos besoins, les programmes utilisés comportent quelques limitations dont il va falloir tenir compte.

Le processus de choix entre les sorties des deux modules

Comme cela a été expliqué précédemment, il faut disposer des scores de reconnaissance pour chaque élément de la phrase gestuelle. Or le système renvoie un score global sur la phrase et non pas sur chaque élément de la phrase. De ce fait, nous n'avons aucun moyen de choisir, à la fin du processus de reconnaissance, la sortie qui nous intéresse. Au sein du prototype implémenté ici, le processus de choix est simulé (voir Paragraphe 3.3.3).

---

4 Programmes utilisés :

buidhmm pour la construction des modèles HMM,
sentrec pour la reconnaissance,
ctx2pic pour déterminer le modèle à partie du signe et de son contexte,
segmerge pour réunir les segments de données par modèle (un fichier par modèle).

Le silence

Les programmes utilisés nécessitent la présence de silences en début et fin de phrase. Notons qu'il n'existe pas de "silence" en geste, puisque les doigts possèdent toujours une valeur de flexion et que la main est toujours située à une position donnée, avec une orientation donnée. Il faut donc choisir un "silence" artificiel. Deux possibilités se présentent : on peut choisir comme configuration-silence une "main au repos" statique ; on peut choisir un geste spécial, qui est peu rencontré dans la langue des signes.

Dans le premier cas, le problème est que la configuration de cette main au repos ressemble beaucoup à la configuration **c** (Figure 3.20). Notre système de mesure ne peut pas les distinguer, comme nous l'avons testé. Cela provoque beaucoup d'erreurs. Il faudrait pouvoir disposer d'un système de mesure permettant de capter la tension musculaire pour distinguer ces deux configurations.

Figure 3.20 : Configurations "*repos*" et **c**.

Dans le deuxième cas, un geste artificiel risque d'être compliqué à réaliser, si l'on veut être sûr qu'il n'est pas utilisé en langue des signes. Cela va induire un biais dans la réalisation du corpus en début et fin de phrase.

Nous avons choisi une solution intermédiaire qui semble donner des résultats satisfaisants. Le "signe-silence" choisi est statique. L'introduction d'un mouvement a semblé vraiment trop artificiel. Il est constitué de la configuration **i** (Figure 3.21). Cette configuration est utilisée pour représenter la lettre "i" en dactylologie. Elle est rarement utilisée parmi les signes (17 signes sur 2526, soit 0,7%). Elle a été choisie plutôt qu'une autre car elle est à la fois assez rarement utilisée et assez simple à exécuter, ce qui limite le biais introduit dans la réalisation du corpus.

Figure 3.21 : Configuration **i**.

La main est orientée paume vers l'arrière et majeur vers la gauche (pour un droitier) et l'emplacement choisi est le ventre du signeur qui est assis, ce qui représente vraiment une

position de repos. Notons que la primitive de mouvement pour le "signe-silence" choisi est l'immobilité (statique).

Ce "signe-silence" a été intégré au corpus présenté dans le paragraphe suivant. Il est utilisé au début et à la fin de chaque phrase.

3.3.2. LE CORPUS

Afin d'évaluer les deux modules du système de reconnaissance, le corpus a été choisi de manière à couvrir plusieurs types de signes : standard, standard variable, classificateur, verbe directionnel. Le sens complet des signes variables ne peut être connu en sortie du processus de reconnaissance car leur interprétation dépend du contexte. Traiter ces signes de manière isolée présente peu d'intérêt. Des structures de phrases bien définies ont été choisies, au sein desquelles ces signes prennent tout leur sens. Notre corpus est ainsi constitué de phrases en LSF, composées de gestes enchaînés.

La difficulté inhérente à l'utilisation d'un système de reconnaissance à base de HMM est l'élaboration de corpus suffisamment vastes et représentatifs du vocabulaire. Chaque geste doit être représenté par un grand nombre d'exemples afin que les variations temporelles et spatiales puissent être correctement exprimées. Comme il n'existe pas de corpus de gestes comme c'est le cas dans d'autres domaines (parole, images), tout un travail doit être fait afin de réaliser un tel corpus.

Les problèmes techniques de capture de gestes semblent à ce jour encore trop importants pour pouvoir élaborer un tel corpus dans de bonnes conditions :

- Le Paragraphe 3.2 indique les limitations importantes du DataGlove. D'une manière générale, les capteurs de gestes ne permettent pas encore une saisie suffisamment fiable et riche comme c'est le cas par exemple pour les microphones en parole.
- Même si l'on disposait d'un système performant pour mesurer les gestes des deux mains, les langues des signes sont aussi constituées de gestes du corps, du visage et même de quelques émissions sinon sonores du moins buccales. De plus, des contacts entre les mains et différentes parties du corps sont souvent effectués (voir Chapitre 2). Nous ne disposons pas à ce jour de capteurs permettant de mesurer toutes les informations utiles.

Lorsque des progrès auront été réalisés dans le domaine des capteurs, certaines règles devront être respectées afin de garantir la généralité du corpus et sa validité :

- Il sera nécessaire de définir un format des données qui soit indépendant des capteurs utilisés. Le plus général serait sans doute d'utiliser un modèle anthropomorphique dans lequel les articulateurs sont représentés [Braffort A., Collet C. et al. 1994a], [Braffort A., Collet C. et al. 1994b].
- Il sera souhaitable que les corpus de langues des signes soient réalisés par des personnes dont c'est la langue maternelle, ce qui nécessitera de mettre en place une collaboration avec les associations dédiées à l'étude ou à l'enseignement des langues des signes.

Comme on peut le voir, l'élaboration d'un tel corpus n'est pas une mince affaire et trouvera son intérêt dans le cadre de développement d'applications d'éducation ou d'outils spécifiques.

Dans le cadre de cette thèse, le but fixé est de proposer un système permettant de traiter des signes dont certains paramètres peuvent être variables en fonction du contexte. Nos expérimentations ont pour rôle principal de mettre à jour tous les problèmes spécifiques à ce type de corpus. Ainsi, nous n'avons pas cherché à concevoir un corpus portant sur une grande quantité de signes. Nous avons surtout cherché à élaborer un corpus de phrases gestuelles dont les structures vont permettre de tester et valider le système de reconnaissance, ainsi que le système de compréhension présenté au Chapitre 4.

Dans ce but, le corpus est constitué de phrases composées de quatre signes, auxquelles deux silences ont été adjoints, au début et à la fin de chacune d'elles. Ces phrases sont construites à partir de sept signes et basées sur cinq structures différentes.

Signes choisis

Les sept signes choisis se répartissent en fonction de la variabilité de leurs paramètres en fonction du contexte. Ils sont illustrés Figure 3.22.{1,2,3} et sont décrits dans le prochain tableau.

Figure 3.22.1 : Signes [**garçon**], [**gâteau**] et [**personne**].

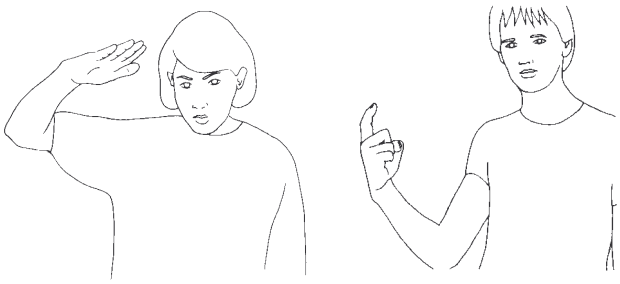


Figure 3.22.2 : Classificateurs [**taille**] et [**index**]

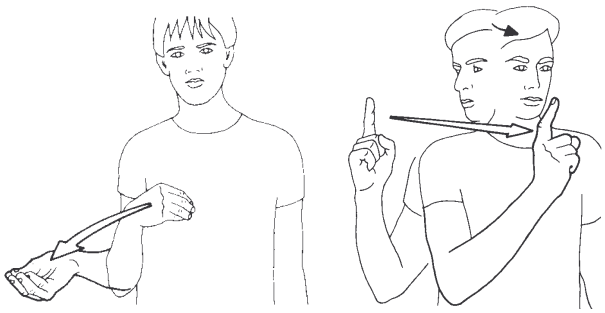


Figure 3.22.3 : Verbes directionnels [**donner**] et [**aller vers**]

- Les deux premiers signes ([**garçon**], [**gâteau**]) sont des signes standard dont les quatre paramètres sont invariables quelque soit le contexte.
- Le signe [**personne**] est un signe standard dont le paramètre d'emplacement est variable : le signeur peut placer une personne dans la scène de narration directement en effectuant le signe à un endroit précis.
- Les signes [**taille**] et [**index**] sont des classificateurs. Seul leur paramètre de configuration et leur primitive de mouvement est invariable. Le premier permet d'indiquer une taille et un emplacement, tandis que le second permet d'indiquer un emplacement et une orientation.
- Les signes [**donner**] et [**aller vers**] sont des verbes directionnels. Seule leur primitive de mouvement est invariable.

Le tableau suivant résume les caractéristiques des signes du corpus :

Nom	Type	Config.	Mouv.	Orient.	Emplac.
garçon	standard	invariable	invariable	invariable	invariable
gâteau	standard	invariable	invariable	invariable	invariable
personne	standard variable	invariable	primitive invariable	variable	variable
taille	classificateur	invariable	primitive invariable	variable	variable
index	classificateur	invariable	primitive invariable	variable	variable
donner	verbe directionnel	variable	primitive invariable	variable	variable
aller vers	verbe directionnel	variable	primitive invariable	variable	variable

Tableau 3.4 : Caractéristiques des signes du corpus.

Structures choisies

Toutes les phrases comportent quatre signes (mis à part les "signes-silences"). Cinq structures de phrases ont été définies dans lesquelles les gestes ont des rôles différents et sept fonctions syntaxiques, qui sont détaillées ci-dessous. Pour chacune d'entre elles, les signes du corpus qui peuvent posséder cette fonction sont indiqués.

- agent ou patient[5]: [garçon|personne|gâteau]
- agent ou patient localisant : personne
- localisant : "index" (classificateur "index" localisant pour objet vertical)
- adjectif qualificatif localisant : taille

5 L'agent est l'être ou la chose qui fait l'action et le patient est l'être ou la chose qui subit l'action [Dubois J., Giacomo M. et al. 1973].

- verbe transitif à un complément : aller vers X
- verbe transitif à deux compléments : donner X à Y

Notons que certaines fonctions sont spécifiques à la langue des signes, notamment celles qui font intervenir la notion de localisation d'un être ou d'une chose dans la scène du signeur. Tous les signes localisants sont exécutés à un emplacement défini dans la scène du signeur.

Deux verbes sont utilisés. Le premier, "donner", implique la présence d'un agent, d'un patient et d'un complément, donc de deux entités animées et d'une entité inanimée. Le deuxième, "aller vers", implique la présence d'un agent et d'un patient, donc de deux entités animées.

Le tableau suivant comporte la description des cinq structures de phrases. Pour chaque signe, la ou les fonctions possibles sont indiquées.

Rôles	*1er geste*	*2ème geste*	*3ème geste*	*4ème geste*
1ère structure	agent ou patient	localisant	patient ou agent	verbe transitif à 2 compléments
2ème structure	agent ou patient	qualificatif localisant	patient ou agent	verbe transitif à 2 compléments
3ème structure	agent ou patient localisant	patient ou agent localisant	patient ou agent	verbe transitif à 2 compléments
4ème structure	agent ou patient	localisant	patient ou agent localisant	verbe transitif à 1 complément
5ème structure	agent ou patient	qualificatif localisant	patient ou agent localisant	verbe transitif à 1 complément

Tableau 3.5 : Description des cinq structures de phrases.

Notons que pour une même structure et une même conjugaison du verbe, l'agent peut précéder ou suivre le patient. L'ordre n'est pas figé, mais certaines séquences se rencontrent plus souvent que d'autres.

Liste des phrases

Pour lever le problème de la variabilité de l'ordre des signes pour une même phrase, les structures syntaxiques choisies sont celles qui correspondent à des signes placés dans l'ordre

le plus fréquent : l'agent est signé avant le patient [Cuxac C. 1987]. Notons que le signeur fait partie intégrante de la scène et est sous-entendu. Il n'est pas nécessaire de signer l'équivalent de "je" ou "moi".

Pour chaque structure, une description générale, la liste des phrases correspondantes ainsi que leur traduction en français sont données. Ces phrases se différencient par l'intermédiaire des signes variables, pour lesquels plusieurs "instanciations" ont été choisies.

♦*Première structure*

Dans cette structure le garçon peut être localisé à deux endroits **loc1** (devant le signeur) et **loc2** (à droite du signeur) à l'aide du classificateur [**index**] et le verbe [**donner**] peut être conjugué à la première ou à la troisième personne du singulier de l'indicatif présent. Quatre phrases gestuelles sont possibles :

[garçon	loc1	gâteau	je-lui-donne]	: "Je donne un gâteau au garçon devant moi"
[garçon	loc2	gâteau	je-lui-donne]	: "Je donne un gâteau au garçon à ma droite"
[garçon	loc1	gâteau	il-me-donne]	: "Le garçon devant moi me donne un gâteau"
[garçon	loc2	gâteau	il-me-donne]	: " Le garçon à ma droite me donne un gâteau"

♦*Deuxième structure*

Ici, pour la taille et l'emplacement d'un individu ([**garçon**] ou [**personne**]), exprimés à l'aide du classificateur [**taille**], quatre choix sont possibles (petit et devant, petit et à droite, grand et devant, grand et à droite) et le verbe [**donner**] peut être conjugué à la première ou à la troisième personne du singulier de l'indicatif présent. Seize phrases sont possibles.

[garçon	taille1+loc1	gâteau	je-lui-donne]	: "Je donne un gâteau au petit garçon devant moi"
[garçon	taille1+loc2	gâteau	je-lui-donne]	: "Je donne un gâteau au petit garçon à ma droite"
[garçon	taille1+loc1	gâteau	il-me-donne]	: "Le petit garçon devant moi me donne un gâteau"
[garçon	taille1+loc2	gâteau	il-me-donne]	: "Le petit garçon à ma droite me donne un gâteau"
[garçon	taille2+loc1	gâteau	je-lui-donne]	: "Je donne un gâteau au grand garçon devant moi"
[garçon	taille2+loc2	gâteau	je-lui-donne]	: "Je donne un gâteau au grand garçon à ma droite"
[garçon	taille2+loc1	gâteau	il-me-donne]	: "Le grand garçon devant moi me donne un gâteau"
[garçon	taille2+loc2	gâteau	il-me-donne]	: "Le grand garçon à ma droite me donne un gâteau"
[personne	taille1+loc1	gâteau	je-lui-donne]	: "Je donne un gâteau à la petite personne devant moi"
[personne	taille1+loc2	gâteau	je-lui-donne]	: "Je donne un gâteau à la petite personne à ma droite"
[personne	taille1+loc1	gâteau	il-me-donne]	: "La petite personne devant moi me donne un gâteau"
[personne	taille1+loc2	gâteau	il-me-donne]	: "La petite personne à ma droite me donne un gâteau"

[personne taille2+loc1 gâteau je-lui-donne] : "Je donne un gâteau à la grande personne devant moi"
[personne taille2+loc2 gâteau je-lui-donne] : "Je donne un gâteau à la grande personne à ma droite"
[personne taille2+loc1 gâteau il-me-donne] : "La grande personne devant moi me donne un gâteau"
[personne taille2+loc2 gâteau il-me-donne] : "La grande personne à ma droite me donne un gâteau"

♦*Troisième structure*

Pour la troisième structure, deux personnes sont localisées directement, la première à **loc1** (devant le signeur) et la deuxième à **loc2** (à droite du signeur). Le verbe [**donner**] peut être conjugué à l'indicatif présent selon six possibilités (combinaison de l'emplacement du signeur, de **loc1** et de **loc2**). Six phrases sont possibles.

[personne+loc1 personne+loc2 gâteau je-donne-a-1] :
"Je donne un gâteau à la personne devant moi et une personne est à ma droite"

[personne+loc1 personne+loc2 gâteau je-donne-a-2] :
"Je donne un gâteau à la personne à ma droite et une personne est devant moi"

[personne+loc1 personne+loc2 gâteau 1-me-donne] :
"La personne devant moi me donne un gâteau et une personne est à ma droite"

[personne+loc1 personne+loc2 gâteau 1-donner-a-2] :
"La personne devant moi donne un gâteau à la personne à ma droite"

[personne+loc1 personne+loc2 gâteau 2-me-donne] :
"La personne à ma droite me donne un gâteau et une personne est devant moi"

[personne+loc1 personne+loc2 gâteau 2-donner-a-1] :
"La personne à ma droite donne un gâteau à la personne devant moi"

♦*Quatrième structure*

Ici, un [**garçon**] est localisé à l'endroit **loc1** (devant le signeur) à l'aide du classificateur [**index**], une [**personne**] est localisée directement à l'endroit **loc2** (à droite du signeur) et le verbe [**aller vers**] peut être conjugué à l'indicatif présent selon six possibilités (combinaison de l'emplacement du signeur, de **loc1** et de **loc2**). Six phrases sont possibles.

[garçon loc1 personne+loc2 je->1] : "Je vais vers le garçon devant moi et une personne est à ma droite"
[garçon loc1 personne+loc2 je->2] : "Je vais vers une personne à ma droite et un garçon est devant moi"
[garçon loc1 personne+loc2 1->moi] : "Le garçon devant moi va vers moi et une personne est à ma droite"
[garçon loc1 personne+loc2 1->2] : "Le garçon devant moi va vers la personne qui est à ma droite"
[garçon loc1 personne+loc2 2->moi] : "La personne à ma droite va vers moi et un garçon est devant moi"

[garçon loc1 personne+loc2 2->1] : "La personne à ma droite va vers le garçon qui est devant moi"

♦*Cinquième structure*

Dans cette structure, deux personnes sont localisées et on montre la taille de la première. Douze phrases sont possibles.

[garçon taille+loc1 personne+loc2 je->1] :
"Je vais vers le petit garçon devant moi et une personne est à ma droite"

[garçon taille+loc1 personne+loc2 je->2] :
"Je vais vers la personne à ma droite et un petit garçon est devant moi"

[garçon taille+loc1 personne+loc2 1->moi] :
"Le petit garçon devant moi va vers moi et une personne est à ma droite"

[garçon taille+loc1 personne+loc2 1->2] :
"Le petit garçon devant moi va vers la personne à ma droite"

[garçon taille+loc1 personne+loc2 2->moi] :
"La personne à ma droite va vers moi et un petit garçon est devant moi"

[garçon taille+loc1 personne+loc2 2->1] :
"La personne à ma droite va vers le petit garçon devant moi"

[personne taille+loc1 personne+loc2 je->1] :
"Je vais vers la petite personne devant moi et une personne est à ma droite"

[personne taille+loc1 personne+loc2 je->2] :
"Je vais vers la personne à ma droite et une petite personne est devant moi"

[personne taille+loc1 personne+loc2 1->moi] :
"La petite personne devant moi va vers moi et une personne est à ma droite"

[personne taille+loc1 personne+loc2 1->2] :
"La petite personne devant moi va vers la personne à ma droite"

[personne taille+loc1 personne+loc2 2->moi] :
"La personne à ma droite va vers moi et une petite personne est devant moi"

[personne taille+loc1 personne+loc2 2->1] :
"La personne à ma droite va vers la petite personne devant moi"

Saisie du corpus

Le corpus a été saisi à l'aide du programme SaiCoS. Il a été réalisé par une seule personne. Notons que l'auteur ayant une main trop petite pour le gant, il a fallu choisir une personne[6] ne connaissant que peu la LSF mais ayant une main de la taille adéquate. La petite taille du vocabulaire a permis un apprentissage rapide des différentes phrases par le signeur. L'important est que les phrases gestuelles soient réalisées de manière "naturelle". L'ordre des phrases a été modifié à chaque saisie, de manière à ce que le corpus ne soit pas appris par coeur et que la variabilité intra-personne du signal soit au mieux représentée.

3.3.3. MISE EN OEUVRE DES HMM

La mise en oeuvre des HMM nécessite le choix de la stratégie de segmentation du corpus d'apprentissage, ainsi que le choix du nombre d'état de chacun des HMM permettant de représenté chacun des signes.

Segmentation

Comme nous cherchons à reconnaître des séquences de gestes, la coarticulation doit être apprise par le système de reconnaissance, au même titre que les signes. Il est possible d'apprendre les gestes de coarticulation indépendamment des gestes du vocabulaire. Cependant, les gestes seront plus facilement différenciés si les coarticulations y sont intégrées, car le geste est alors représenté en contexte et non pas isolément (de la même manière, dans le domaine de la parole, les phones sont modélisés en contexte).

Pour segmenter les différentes parties des gestes, une procédure semi-automatique a été réalisée. Elle est composée de trois étapes :

- Pré-segmentation manuelle.

  Le début et la fin de chaque signe sont saisis manuellement.

  La segmentation du corpus d'apprentissage est réalisée à l'aide de l'outil VAG2, à partir d'une observation visuelle de la main fil de fer et des courbes de valeurs (13 courbes de base sont disponibles : les dix valeurs de flexion des doigts et les coordonnées x, y et z représentant la position - on peut y ajouter les dérivée première et seconde).

6 Mes plus grands remerciements à Thierry Lebourque pour ses "coups de main" !

- Segmentation automatique des gestes de coarticulation.

  Chaque geste de coarticulation est divisé en deux parties égales. Leur milieu est calculé automatiquement à partie de la segmentation manuelle précédente.
- Regroupement automatique des différentes parties constituant chaque signe. Les moitiés de gestes de coarticulations précédent et suivant sont intégrés à chaque signe.

Le processus de segmentation du corpus d'apprentissage est illustré dans la Figure 3.23, qui représente une phrase composée de quatre signes précédés et suivis de silences. La courbe donnée dans cette Figure n'a pas de sens réel. Elle n'est là que pour illustrer le processus.

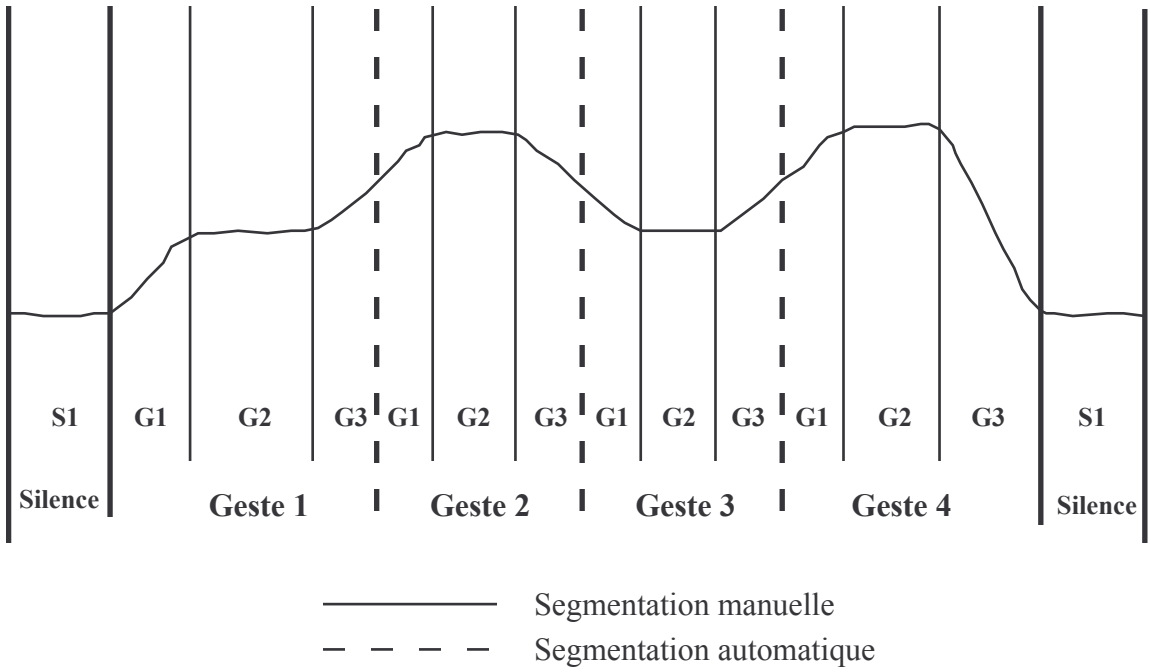


Figure 3.23 : Processus de segmentation.

Ainsi, un signe est composé de trois parties :

- **G1**, la moitié de la transition entre le signe courant et le signe précédent,
- **G2**, la partie stable du signe courant,
- **G3**, la moitié de la transition entre le signe courant et le signe suivant.

Ces trois parties représentent la structure du signe. L'aspect structurel des HMM permet de représenter la structure interne des données à l'aide des états. Dans le cas présent, les trois éléments structurels donnent lieu à trois états.

Le corpus choisi ne contient que des signes dont la structure interne est simple (pas de variation importante des paramètres dans la partie **G2**). L'introduction de nouveaux signes nécessitera l'étude de leur structure interne afin de vérifier si cette partie **G2** n'est pas décomposable en plusieurs parties. Ce pourrait être le cas pour un signe possédant une répétition, tel que [**éponge**] par exemple (Figure 3.24).

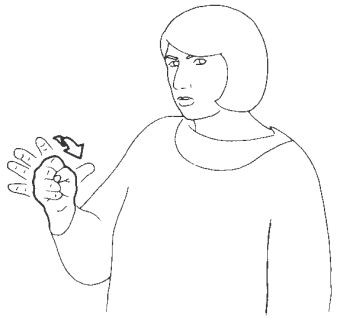


Figure 3.24 : Le signe [**éponge**].

Un cas particulier de signe est le silence. En début de phrase, il n'est pas précédé d'une transition. De même, en fin de phrase, il n'est pas suivi d'une transition. Par simplification, le silence est représenté avec un HMM à un seul état (**S1**). De ce fait, les signes qui suivent un silence sont segmentés de manière à ce que toute la phase de transition avec le silence soit représentée par le premier état du HMM représentant le premier signe (**G1**). De même, les signes qui précèdent un silence sont segmentés de manière à ce que toute la phase de transition avec le silence soit représentée par le dernier état du HMM représentant le dernier signe (**G3**).

Pour un même signe, les première et troisième parties représentant les transitions (**G1**, **G3**) peuvent être différentes si les signes qui précèdent ou suivent ce signe sont différents. Ainsi, pour un même signe, plusieurs triplets peuvent coexister. A chaque triplet correspond un modèle HMM. Les paragraphes suivants contiennent la liste des HMM de chacun des modules de reconnaissance.

Modèles de signes standard

Seuls deux signes possèdent quatre paramètres invariables. Il s'agit des signes [**garçon**] et [**gâteau**]. Comme nous l'avons indiqué, il n'est pas possible avec les outils utilisés de connaître les scores de reconnaissance de chacun des signes. De ce fait, il est difficile d'utiliser l'architecture du système de reconnaissance telle que nous l'avons idéalement

définie. La seule possibilité est d'apprendre l'ensemble du vocabulaire au module dédié aux signes standard et de n'utiliser que sa sortie dans les traitements ultérieurs.

Le vocabulaire utilisé pour l'apprentissage de ce module est donc constitué des sept signes [**garçon**], [**gâteau**], [**personne**], [**taille**], [**localisation**], [**donner**] et [**aller-vers**], auquel le "signe-silence" a été ajouté.

Le tableau suivant contient les différents triplets (**G1**, **G2**, **G3**), ainsi que leur nombre d'occurrences au sein du corpus d'apprentissage. La notation utilisée pour la première et la troisième partie correspond respectivement aux signes précédent et suivant. Le symbole — indique le "signe-silence".

Signes	Précédent	Courant	Suivant	Occurrences
garçon	—	garçon	localisation	10
	—	garçon	taille	14
gâteau	localisation	gâteau	donner	4
	taille	gâteau	donner	16
	personne	gâteau	donner	6
personne	—	personne	taille	14
	—	personne	personne	6
	personne	personne	gâteau	6
	localisation	personne	aller-vers	6
	taille	personne	aller-vers	12
taille	garçon	taille	gâteau	6
	personne	taille	gâteau	6
	garçon	taille	personne	6
	personne	taille	personne	6
localisation	garçon	localisation	gâteau	4
	garçon	localisation	personne	6
donner	gâteau	donner	—	26
aller-vers	personne	aller-vers	—	18

Tableau 3.6 : Triplets du corpus pour le module des signes standard.

Si suffisamment d'exemples sont saisis pour chaque signe, 18 HMM différents peuvent être construits pour le module dédié aux signes standard.

Modèles de signes variables

Six configurations différentes sont dénombrées au sein du vocabulaire choisi, qui sont **1**, **s**, **sc**, **5**, **index**, **bec5**, ainsi que celle utilisée pour le silence **i**. Elles sont illustrées Figure 3.25. Notons que ces six configurations "couvrent" 29% du corpus étudié au Chapitre 2 (et même 44%, si l'on considère que le capteur utilisé ne différencie pas les configurations 5-plat-moufle et bec5-angle90).

Figure 3.25 : Les configurations rencontrées dans le corpus.

Seules les deux primitives de mouvement "statique" et "droite" sont présentes dans le corpus. En réalité, ce qui parait être une droite dans le dictionnaire est plutôt un arc, du fait des rotations induites par les articulations, comme cela a pu être observé à l'aide de l'outil d'analyse du mouvement (TePa). Pour simplifier, ces deux primitives ont été intitulées "mobile" et "immobile".

Le vocabulaire du module dédié aux signes variables est constitué de toutes les combinaisons des configurations et des primitives présentes dans le corpus. La seule configuration pour laquelle on rencontre une primitive mobile et immobile est **index**. Nous avons spécifié le vocabulaire en reprenant le code utilisé pour les configurations et en y ajoutant un **i** ou un **m** selon que la primitive est "immobile" ou "mobile".

Les différents triplets présents dans le corpus sont indiqués dans le tableau suivant .

Configuration	Précédent	Courant	Suivant	Occurrences
1-m	—	1-m	ind-i	10
	—	1-m	5-i	14
s-m	ind-i	s-m	bec5-m	4
	5-i	s-m	bec5-m	16
	sc-m	s-m	bec5-m	6
sc-m	—	sc-m	5-i	14
	—	sc-m	sc-m	6
	sc-m	sc-m	5-i	6
	5-i	sc-m	ind-m	12
	ind-i	sc-m	ind-m	6
5-i	1-m	5-i	s-m	8
	sc-m	5-i	s-m	8
	1-m	5-i	sc-m	6
	sc-m	5-i	sc-m	6
ind-i	1-m	ind-i	s-m	4
	1-m	ind-i	sc-m	6
ind-m	sc-m	ind-m	—	18
bec5-m	s-m	bec5-m	—	26

Tableau 3.7 : Triplets du corpus pour le module des signes variables.

Ici encore, si suffisamment d'exemples sont saisis, 18 HMM différents peuvent être construits pour le module dédié aux signes variables.

3.3.4. PREMIERS RESULTATS

Des taux de reconnaissance très encourageants ont été obtenus pour le module dédié aux signes standard (étendu à tous les signes du corpus comme cela a été expliqué précédemment) et le module dédié aux signes variables. Les résultats sont présentés dans le tableau suivant. Ils ont été obtenus à partir d'un corpus constitué de deux ensembles de 44 phrases, le premier pour l'apprentissage et le second pour la reconnaissance.

Module	Corrects	Erreurs
Signes standard	96%	4%
Signes variables	92%	8%

Tableau 3.8 : Taux de reconnaissance.

Le détail des taux d'erreurs sont calculés automatiquement par un des programmes constituant l'ensemble des outils de mise en oeuvre des HMM.

Module	Substitution	Suppression	Insertion
Signes standard	0%	0,5%	3,5%
Signes variables	0%	0,5%	7,5%

Tableau 3.9 : Détail des taux d'erreurs.

Les différents types d'erreurs possibles sont des substitutions, des suppressions ou des insertions.

- Dans le cadre du corpus utilisé, aucune erreur de substitution n'a été observée. Les signes qui forment ce corpus sont sans doute suffisamment distincts.
- Les erreurs de type suppression sont peu nombreuses (une seule pour chacun des modules) et concernent le signe [**personne**] qui est répété ([**personne**] [**personne**] [**gâteau**] [**donner**]). La différence entre les deux signes [**personne**] se situe dans les valeurs d'emplacement x, y, z et de mouvement Δx, Δy et Δz.
- Les erreurs de type insertion sont les plus nombreuses. Elles apparaissent surtout lorsque le verbe [**aller**] est employé. Les signes insérés sont alors [**localisation**] et

[**aller**]. Ces signes diffèrent uniquement dans les valeurs d'emplacement x, y, z et de mouvement Δx, Δy et Δz.

Comme on pouvait s'y attendre, ces erreurs sont liées au fait que le paramètre de mouvement n'a pas le même poids que les autres paramètres. Ce problème est encore plus présent dans le module dédié aux signes variables.

L'utilisation d'un outil dans lequel il sera possible de spécifier le poids relatif des données permettra d'améliorer sensiblement les performances de notre système.

3.4. CONCLUSION

Dans ce chapitre, nous avons présenté un prototype de système de reconnaissance de gestes de la LSF capable de traiter les deux catégories de signes distinguées : les signes standard, dont les paramètres sont indépendants du contexte, et les signes variables, pour lesquels au moins un paramètre varie en fonction du contexte. La technique de reconnaissance utilisée est basée sur la mise en oeuvre de modèles de Markov cachés, qui ont permis l'obtention de taux de reconnaissance très encourageants, sur les deux catégories de signes (respectivement 96% et 92%). Les erreurs du système de reconnaissance proviennent généralement du fait que les paramètres articulatoires constituant le vecteur de données n'ont pas le même poids.

Ce premier prototype pourra évoluer vers un système plus performant lorsque l'on disposera d'une version de l'outil HMM permettant de séparer le vecteur de paramètres en flux de données afin de donner un poids identique à chacun de ces flux (cette évolution est envisagée aussi dans le domaine de la parole, pour mettre en oeuvre l'intégration des reconnaissances phonémique et prosodique [Geoffrois E. 1995]). Lorsque l'on disposera d'un système permettant de capter les gestes des deux mains, la taille du vocabulaire pourra considérablement s'étendre, ainsi que les types de signes traités.

Les performances du prototype sont cependant suffisantes pour envisager l'étude du système de compréhension qui doit être raccordé à la sortie du système de reconnaissance. Son but est de compléter l'interprétation des signes variables. Cette étude est présentée dans le chapitre suivant.

Chapitre 4

VERS UN SYSTEME DE COMPREHENSION

L'objet de ce chapitre est de proposer un système de compréhension de phrases de la LSF, connecté au système de reconnaissance décrit dans le chapitre précédent.

Le premier objectif est de compléter les informations fournies par le module de reconnaissance. En effet, les signes de type standard reconnus sont remplacés par un symbole les représentant, mais pour les signes de type variable, les symboles fournis ne concernent que la configuration et la primitive de mouvement. Il est nécessaire de reconnaître quelle est la fonction syntaxique du signe afin de récupérer les valeurs d'emplacement, d'orientation ou de direction du mouvement, pour que l'interprétation du signe soit complète.

Le deuxième objectif est de proposer un modèle de représentation de la phrase gestuelle permettant par la suite de la traduire en une phrase en français.

Ce chapitre comporte quatre parties. La première décrit le modèle permettant de représenter les entités présentes dans le discours, ainsi que leur relations spatiales. La deuxième présente l'architecture du système de compréhension de phrases en LSF. La troisième partie illustre le fonctionnement du système sur un exemple. Enfin, la dernière partie indique les applications possibles d'un tel système.

4.1. MODELISATION DE LA SCENE DE NARRATION

L'utilisation d'une grammaire permet d'améliorer efficacement les performances des systèmes de reconnaissance (voir Chapitre 3). Toutefois, comme l'ordre des signes est beaucoup moins significatif en LSF que l'arrangement spatial des gestes entre eux, les grammaires statistiques (basées sur des fréquences de succession de symboles) ne sont pas suffisantes si l'on veut traiter des phrases dans lesquelles les informations spatiales sont pleinement utilisées. Il faut développer un autre type de grammaire, qui prenne en compte la spatialité de la syntaxe de la LSF et la structure des signes.

Le système de reconnaissance permet de classifier les signes standard ainsi que le couple {configuration, primitive de la trajectoire du mouvement} pour tous les signes. Le paramètre de configuration et la primitive de la trajectoire du paramètre mouvement sont associés à un symbole les représentant. En revanche, les paramètres Orientation et Emplacement ne possèdent pas de symbole permettant de les représenter. Et dans certains cas, leurs valeurs numériques sont nécessaires au niveau de l'interprétation.

Le message que veut transmettre un signeur n'est pas uniquement une séquence de signes, c'est plutôt la description d'une image en trois dimensions qui est effectuée. La compréhension du message par l'interlocuteur passe par la reconstruction de cette scène de narration dans l'espace. Pour interpréter le message après la phase de reconnaissance, il faut mettre en oeuvre une représentation informatique de la scène de narration du signeur.

Cette représentation doit inclure les entités, animées ou inanimées, présentes dans le discours.

4.1.1. REPRESENTATION DES ENTITES

Pour représenter une entité, il est nécessaire de connaître sa forme, son emplacement, son orientation, son mouvement éventuel, ainsi que les relations spatiales avec les autres entités.

4.2.1.1. Forme d'une entité

Pour décrire la forme d'une entité en LSF, plusieurs stratégies sont possibles. Certaines des stratégies font intervenir des signes nommés *descripteurs*, qui sont utilisés dans des procédés de narration nommés *transferts*. Elles font souvent intervenir différentes parties du

corps (rotation du buste, mouvement des épaules…). Ne disposant pas de moyen de capter ces mouvements à ce jour, nous ne les avons pas incluses dans notre étude. Pour plus de détails, voir [Cuxac C. 1987] et [Cuxac C. 1993a].

La manière la plus simple de décrire une forme consiste à représenter un trait saillant de l'entité. Ce trait saillant correspond en général à un volume, tracé à l'aide de la configuration de la main. Il est possible d'y adjoindre un mouvement afin de décrire plus précisément une particularité de la forme (exemple, le signe [**armoire**] Figure 2.36) ou une fonctionnalité (exemple, le signe [**éponge**] Figure 3.24).

Quand il est nécessaire d'indiquer une partie d'une entité, animée ou inanimée, une description hiérarchique est utilisée. Par exemple, si le signeur veut montrer une cheminée placée sur le toit d'une maison , il commence par décrire la maison (Figure 4.1 : 1g - 1d), puis, avec la main dominée il reprend le toit (Figure 4.1 : 2g), tandis qu'avec la main dominante il représente la cheminée (Figure 4.1 : 2d), en plaçant cette main sur la main dominée. S'il veut représenter la fumée sortant de la cheminée, il procède de manière similaire : la main dominée reprend la cheminée (Figure 4.1 : 3g) tandis que la main dominante représente la fumée sortant de la cheminée (Figure 4.1 : 3d). Dans cet exemple, on suppose que le signeur est droitier. Sa main dominée est alors la main gauche (g), tandis que la main droite est la main dominante (d).

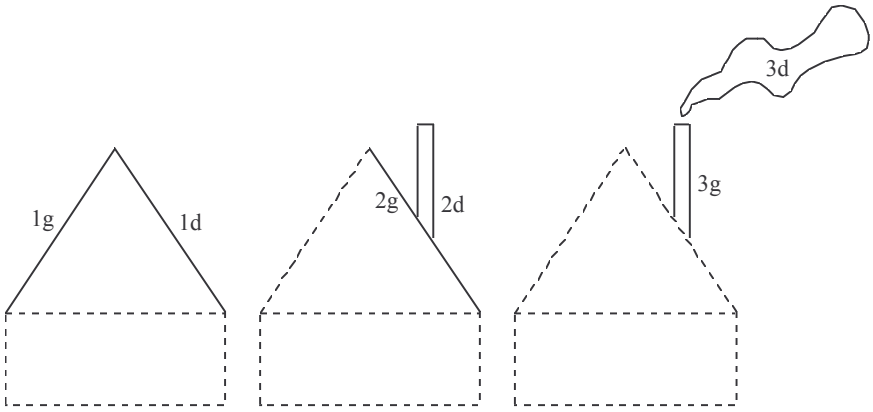


Figure 4.1 : Exemple de description d'un objet complexe (trois étapes).

La stratégie de description est hiérarchique. La partie la plus générale est décrite en premier. Puis l'attention est portée sur des parties de plus en plus précises, avec autant d'étapes qu'il est nécessaire pour pouvoir communiquer avec précision l'emplacement de la

partie dont il est question par rapport à l'entité de base. Chaque nouvelle partie de l'entité est décrite et placée en fonction de la partie précédemment décrite.

Si l'on se réfère aux travaux de David Marr [Marr D. 1982] qui font référence dans le domaine de la représentation de la forme des objets, trois aspects doivent être pris en compte pour choisir une représentation :

- le système de coordonnées, qui peut être :
  - centré sur celui qui voit l'entité,
  - ou centré sur l'entité.
- les primitives, qui peuvent être :
  - surfaciques,
  - ou volumiques.
- l'organisation, qui peut être :
  - non structurée,
  - ou structurée.

Si l'on veut pouvoir décrire une entité à la fois d'une manière grossière et à la fois d'une manière plus détaillée, le type de représentation correspondant doit posséder un système de coordonnées centré sur l'entité, avec des primitives volumiques et une organisation structurée. Ce type de représentation est illustré dans la Figure 4.2, extraite de [Marr D. 1982].

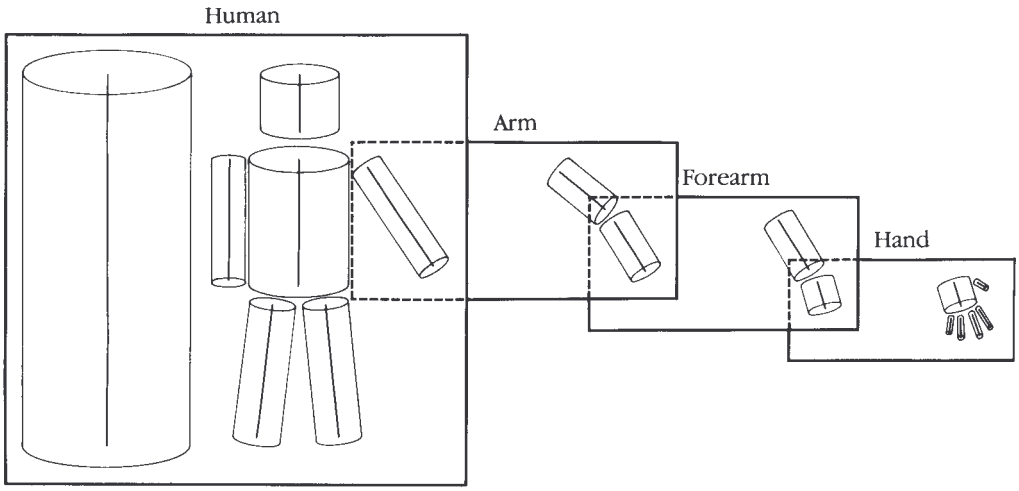


Figure 4.2 : Représentation d'une personne [Marr D. 1982].

Pour représenter les différentes entités dans la scène de narration, il est nécessaire de développer une base de connaissance sur ces entités. Cette base comporte en particulier un module dans lequel la forme des entités est stockée selon le formalisme décrit ci-dessus.

4.2.1.2. Emplacement et orientation des entités

Avant l'étape de compréhension des phrases de LSF, un processus de reconnaissance doit être effectué. Ce processus convertit les valeurs numériques données par un système de capture (ici, le gant numérique) en valeurs symboliques. Pour les signes variables, cela concerne la configuration et la primitive de mouvement. En parallèle, pour représenter la scène de narration, il est nécessaire de mémoriser les valeurs numériques d'emplacement (x, y, z) et d'orientation (α, β, γ).

En général, les mesures de l'emplacement et de l'orientation sont absolues (x_a, y_a, z_a, et α_a, β_a, γ_a). C'est le cas pour notre gant numérique. Le repère est fixe et placé devant le signeur. Or, quand le signeur construit son message, il place les entités intervenant dans le discours en fonction de sa propre position. Pour la scène de narration, il faut utiliser des valeurs d'emplacement et d'orientation relatives à un système de coordonnées placé sur le signeur (x_S, y_S, z_S, et α_S, β_S, γ_S). Il suffit pour cela de connaître l'emplacement et l'orientation absolus du signeur et d'utiliser des matrices de transformation (translation, rotation) pour passer d'un repère à l'autre.

Ainsi, lorsqu'il est nécessaire de connaître l'emplacement et l'orientation d'une entité du discours, il suffit de récupérer les valeurs brutes issues du gant et de les transcrire dans le repère du signeur afin les situer par rapport au signeur dans la scène de narration.

4.2.1.3. Relations spatiales entre les entités

Quatre groupes de lexèmes spatiaux entrent en jeu pour exprimer des relations spatiales en langage naturel [Briffault X. 1992] :

- Les prépositions spatiales (sur, dans, à droite, devant, entre…)
- Les adverbes spatiaux (loin, près…)
- Les noms et les adjectifs spatiaux (grand, large, debout, extérieur…)
- Les verbes spatiaux (être assis, marcher, trembler…)

Dans les langues des signes, les prépositions et les adverbes spatiaux ne sont pas nécessaires, car les relations spatiales sont explicites dans la scène de narration.

Comme la scène contient l'emplacement et l'orientation des entités par rapport au signeur, il est possible d'en déduire directement les relations spatiales entre les entités. En fait, ces valeurs d'emplacement et d'orientation sont d'une précision relative. Elles n'ont pour but que d'indiquer les relations spatiales.

Lorsqu'il est nécessaire de représenter de manière précise les relations spatiales entre les objets, un moyen simple est d'utiliser des arcs étiquetés, permettant de joindre deux entités dont on veut spécifier les relations spatiales.

4.1.2. REPRESENTATION DE LA SCENE DE NARRATION

La représentation de l'ensemble des entités et de leurs relations spatiales est un graphe comportant autant de noeuds qu'il y a d'entités dans le discours. Ces noeuds contiennent des informations sur la forme, l'emplacement et l'orientation des entités, selon les principes décrits précédemment. Les arcs du graphe contiennent les informations sur les relations spatiales. Quand une forme est complexe, elle peut être décomposée en plusieurs noeuds reliés par des arcs, de manière à obtenir une description hiérarchique de l'entité.

Par exemple, une phrase en LSF signifiant "Il y a un verre sur la table devant moi, avec un glaçon dedans, et il y a un ballon sous la table." est exprimée en construisant une image 3D telle que celle illustrée Figure 4.3.

Figure 4.3 : Exemple de scène de narration.

Elle sera représentée, à l'aide du graphe, de la manière suivante :

- Les entités présentes dans cette phrase sont "verre", "table", "glaçon" et "ballon". Ils correspondent à des signes standard. Les cercles dans la Figure 4.2 représentent les entités.
- Leurs formes peuvent être prédéfinies et stockées dans une base de connaissances.
- Leur emplacement et leur orientation sont indiqués au moyen de classificateurs, qui reprennent un trait saillant dans la forme de l'entité. Ces classificateurs sont réalisés par les deux mains. La main dominée indique l'emplacement d'une entité E1 préalablement décrite, tandis que la main dominante place la nouvelle entité E2 par rapport à E1.

Dans l'exemple cité précédemment, la première entité signée est la table, car c'est l'élément par rapport auquel on peut placer les autres entités. La deuxième entité signée est le verre. Pour le placer SUR la table, la main dominée reprend la forme de la table à l'aide d'un configuration "main-plate", tandis que la main dominante reprend la forme du verre (forme C) et est placée de manière à ce que l'on voit que le verre est SUR la table. De la même manière, on indique que le glaçon est DANS le verre et que le ballon est SOUS la table. La séquence complète est :

[**table**] [**verre**] [*plat*]***sur***[*C*] [**glaçon**] [*C*]***dans***[*bec5*] [**ballon**] [*plat*]***sous***[*boule*]

Les expressions de type [*config1*]***préposition***[*config2*] représentent la configuration config1 de la main dominée et la configuration config2 de la main dominante. Ces deux signes de type classificateur permettent, par le biais du paramètre d'emplacement, d'indiquer la relation spatiale entre les deux entités référencées.

Dans le cas de notre exemple, on obtient le graphe illustré Figure 4.4.

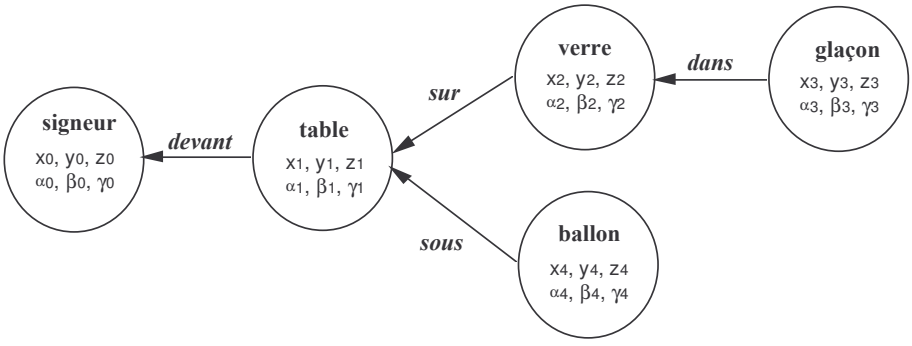


Figure 4.4 : Exemple de graphe.

4.2. ARCHITECTURE DU SYSTEME ARGO

L'architecture générale du système nommé ARGo (Analyse et Reconnaissance de Gestes sémiOtiques) comporte le système de **Reconnaissance** présenté au chapitre précédent et le système de **Compréhension**. Elle est illustré Figure 4.5 :

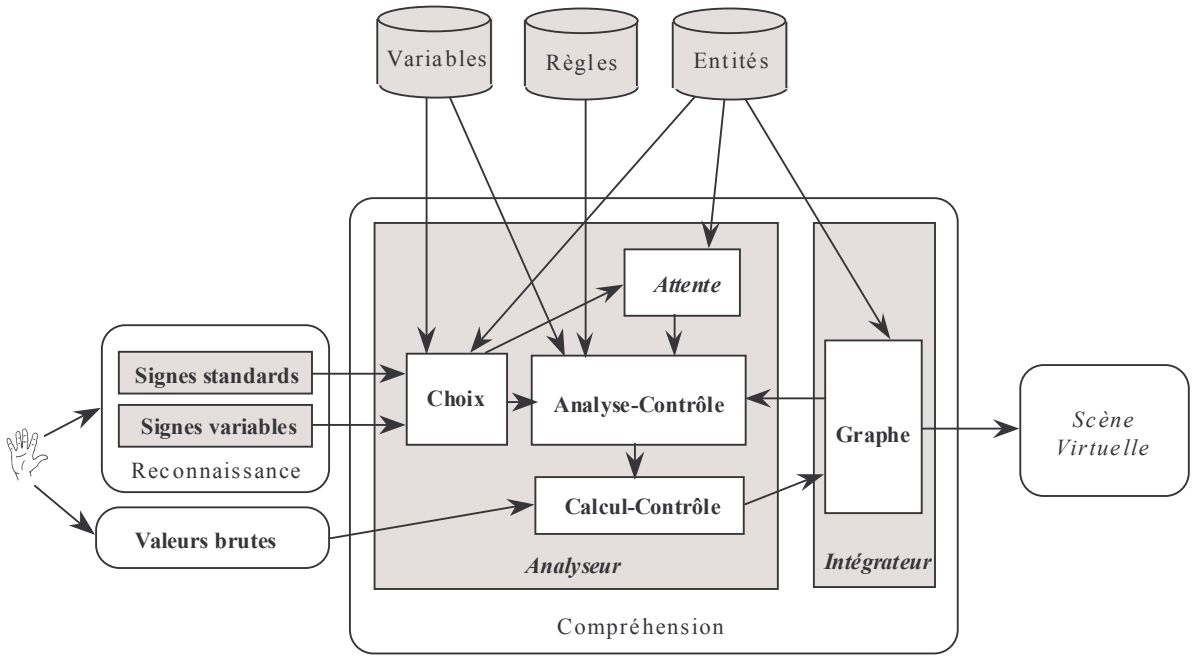


Figure 4.5 : Architecture détaillée du système ARGo.

Le processus de compréhension est effectué en deux étapes. La première, réalisée par un sous-module nommé **Analyseur**, analyse la sortie du module de reconnaissance afin de détecter le type de geste qui a été effectué (signe standard, classificateur, verbe directionnel). Selon le type du signe, l'Analyseur récupère les **Valeurs brutes** d'emplacement et d'orientation (x, y, z et α, β, γ) et calcule au besoin la direction du mouvement (pour les verbes directionnels).

La seconde étape, réalisée par un sous-module nommé **Intégrateur**, réalise l'intégration de toutes les informations produites lors de l'analyse et met à jour le graphe décrit précédemment.

Les différentes parties de l'architecture sont détaillées dans les paragraphes suivants.

4.2.1. ANALYSEUR

L'analyseur récupère les sorties du module de reconnaissance. Ce dernier envoie deux résultats, le premier en provenance du module dédié aux signes standard et le second, du module dédié aux signes variables. Ces résultats sont les deux séquences de symboles S1 et S2 spécifiées au chapitre précédent.

L'analyse se déroule selon trois étapes :

- La première consiste à choisir la sortie du module de reconnaissance parmi les deux sorties fournies. Il s'agit du processus de **Choix**, présenté dès le Chapitre 3.
- La deuxième consiste à effectuer l'analyse fonctionnelle du signe et les contrôles de compatibilité par rapport au contexte. C'est le processus d'**Analyse-Contrôle**.
- La troisième étape, le processus de **Calcul-Contrôle**, consiste à récupérer les valeurs brutes éventuellement utiles afin de calculer les informations nécessaires à l'interprétation complète du signe et de vérifier leurs cohérences.

4.2.1.1. Processus de Choix

La première étape de l'analyse consiste à choisir, pour chaque signe, un des deux résultats, en fonction du score de reconnaissance fourni par chaque sous-module. Rappelons que l'outil de reconnaissance dont nous disposons ne possède pas cette fonctionnalité. Nous ne disposons que d'un score global sur la phrase. Cette étape n'est donc pas implémentée actuellement. Comme cela l'a été expliqué dans le chapitre précédent, elle est simulée. Seule la sortie du sous-module dédié aux signes standard est considérée, car on lui a fait apprendre l'ensemble du vocabulaire (voir Paragraphe 3.3.3 "*Modèles de signes standard*").

Deux bases de données représentant le vocabulaire traité sont utilisées à différents niveau du système de compréhension et en particulier lors du processus de choix :

- La Base de Données des Entités (BDE) comporte toutes les informations relatives aux signes standard.
- La Base de Données des Variables (BDV) comporte toutes les informations relatives aux signes variables.

Elles sont décrites plus précisément au Paragraphe 4.2.3.

A l'aide de ces deux bases de données, le sous-module Choix fournit la catégorie du signe (standard - variable) en comparant le symbole produit par le système de reconnaissance

aux différents symboles contenus dans les deux bases de données. En fonction de cette catégorie, la suite du processus diffère :

- Lorsqu'un signe standard est détecté, il est placé en **Attente**. En effet, il n'est pas possible à ce niveau de l'ajouter au graphe, son emplacement et son orientation n'étant pas connus.
- Lorsqu'un signe variable est détecté, sa fonction syntaxique est récupérée dans la BDV afin de déterminer la règle spatio-temporelle qu'il faut ensuite lui appliquer. Ces règles sont décrites maintenant.

4.2.1.2. Processus d'Analyse-Contrôle

Ce processus permet d'associer à un signe variable les entités correspondantes afin de compléter l'interprétation. Il est basé sur un ensemble de règles de spatio-temporelles, chacune dédiée à un type de signe.

Pour le moment, quatre règles ont été étudiées. Elles concernent les classificateurs, les verbes directionnels, les verbes directionnels incluant un classificateur et les signes déictiques. Ce sont des signes couramment utilisées en LSF et pour lesquelles nos limitations d'ordre technique (un seul système de capture pour main droite uniquement) ne posent pas de problèmes pour l'évaluation : il est possible de trouver des phrases comportant ce type de signes et réalisées uniquement avec la main droite.

Pour simplifier le problème, les règles spatio-temporelles proposées sont basées sur les hypothèses suivantes :

- Un classificateur n'est signé que si l'entité correspondante a été signée au préalable.
- De même, un verbe n'est signé que si les entités mises en jeu ont été signées au préalable.
- Un déictique n'est signé qu'après l'entité qu'il désigne.

Cette logique dans l'ordre des signes correspond à celle qui est le plus souvent employée en langue des signes. L'ordre naturel est Agent-Patient-Action et Localisant-Localisé [Cuxac C. 1987].

Lors d'une prochaine étape d'évolution du système de compréhension, il faudra prendre en compte les cas pour lesquels cet ordre n'est pas respecté. Cela pourra être réalisé sur le même principe que la zone d’attente utilisée pour les signes standard.

Ces règles sont basées sur celles de la LSF [Moody B. 1983].

- **Classificateur**

  Règle LSF :

  Un classificateur est un signe qui décrit et représente toute une classe d'objets par l'intermédiaire de leur forme. Ils ont une fonction de "super-pronom", c'est-à-dire qu'ils sont utilisés pour décrire la forme des entités et pour les localiser dans l'espace. Les paramètres de mouvement, d'orientation et d'emplacement varient selon le contexte. Seule la configuration est indépendante du contexte.

  Processus mis en place :

  Pour interpréter un classificateur, il faut retrouver à quelle entité il fait référence. Pour simplifier, on ne considère ici que les classificateurs dont la fonction est de placer une entité dans la scène de narration. De ce fait les entités qu'il est possible d'associer au classificateur sont stockées dans la zone d'attente. Elles sont récupérées, afin de déterminer la liste des configurations possibles. Des configurations spécifiques sont dédiées à des entités spécifiques (objets plats, ronds, personnes…) (voir Chapitre 2). Le processus renvoie la liste des entités pour lesquelles un des classificateurs possibles est celui en cours d'analyse. Si aucune entité ne correspond, ou si plus d'une entité correspond, un message d'erreur est émis.

  Exemple :

  Si une entité "garçon" a été reconnue et placée en attente et si le signe suivant est un classificateur "index" qui peut être utilisé pour représenter une personne, alors, par comparaison des possibilités, le processus infère que ce classificateur peut se rapporter au garçon et cette entité est renvoyée.

- **Verbes directionnels**

  Règle LSF :

  Les verbes directionnels se conjuguent dans l'espace. La direction du mouvement et l'orientation de la main permettent de déterminer les rôles d'agent et de patient. Seuls la configuration et la primitive de la trajectoire du mouvement sont indépendants de la conjugaison.

  Processus mis en place :

  Les entités présentes dans le graphe et celles placées en attente sont récupérées en vue de déterminer lesquelles il est possible d'associer au verbe en fonction de ses caractéristiques (agent, patient, objet, source, but). Si l'association n'est pas

complète, ou si plus d'une entité correspond dans la zone d'attente, un message d'erreur est émis.

Exemple :

Pour le verbe "donner", il faut qu'il y ait au moins deux entité animées dans le graphe et une entité inanimée, dans le graphe ou en attente, pour que l'interprétation de ce verbe soit possible.

- **Verbes directionnels incluant un classificateur**

Règle LSF :

Parfois, le verbe peut intégrer un classificateur qui jouera le rôle de super-pronom (voit Chapitre 2) et dans ce cas, seule la primitive de mouvement est invariable.

Processus mis en place :

Comme précédemment, les entités présentes dans le graphe et celles mises en attente sont récupérées, pour déterminer l'association entités-verbe. Si cette liste n'est pas complète ou est ambiguë, un message d'erreur est émis.

De plus, pour chacune des entités pouvant être représentées par le classificateur intégré dans le verbe, un contrôle est réalisé sur le paramètre de configuration, de la même manière que précédemment pour les classificateurs. Si un problème d'incompatibilité est détecté, un message d'erreur est émis.

- **Déictiques**

Règle LSF :

Un signe de type déictique est employé pour désigner une personne, un objet ou un événement.

Processus mis en place :

L'entité désignée doit avoir été signée et placée dans la scène de narration avant le geste de désignation, ce qui implique qu'elle est stockée dans le graphe. Les entités présentes dans le graphe sont récupérées afin de déterminer laquelle correspond au signe déictique. Si aucune entité correspondante n'est présente dans le graphe, un message d'erreur est émis.

Le processus d'Analyse-Contrôle permet d'associer à chaque signe variable les entités concernées. S'il n'a pas été possible d'effectuer cette association, l'analyse s'arrête. Sinon, le processus de calcul est déclenché afin de positionner précisément les entités dans le graphe et de compléter les contrôles et l'interprétation.

4.2.1.3. Processus de Calcul-Contrôle

Ce processus permet d'effectuer des contrôles qui portent sur la cohérence des valeurs numériques. En fonction du type de signe variable détecté, les valeurs numériques qu'il est nécessaire de récupérer diffèrent :

- Dans le cas des classificateurs, les valeurs d'emplacement et d'orientation sont recueillies, pour placer l'entité correspondante dans le graphe.
- Dans le cas des verbes directionnels et des verbes directionnels incluant un classificateur, il faut récupérer les emplacements initiaux et finaux du mouvement, pour déterminer l'agent et le patient. S'il n'y a pas correspondance entre ces emplacements et les entités possibles, un message d'erreur est envoyé.
- Dans le cas d'un déictique, il faut recueillir la direction correspondante. Si aucune entité n'est placée dans la zone pointée, un message d'erreur est envoyé.

Notons que pour ces contrôles, ce sont des valeurs numériques qui sont manipulées et non pas des symboles. Des problèmes d'ambiguïté, de précision vont alors se poser. Si l'on se contente d'une approche de type analytique, par comparaison de valeurs avec l'utilisation de seuils fixés a priori, cela risque de provoquer des erreurs à ce niveau. Il faut envisager une approche plus robuste (il serait peut-être possible de passer par un apprentissage et d'utiliser une approche markovienne). Dans l'état actuel de l'implémentation du prototype, le système propose un choix correspondant à l'entité dont l'emplacement est le plus proche, par calcul de distance.

Lorsqu'un signe variable a été analysé et que les valeurs numériques nécessaires ont été calculées, il reste à mettre à jour les différentes composantes de représentation de la phrase. Ce travail est réalisé par l'Intégrateur.

4.2.2. INTEGRATEUR

Ce module met à jour la zone d'attente et le graphe. L'intégrateur fonctionne de manière différente, en fonction du type de signe qui a été reconnu et analysé lors des étapes précédentes :

- **Signe standard**

  L'intégrateur ne fait rien, car toutes les informations ne sont pas encore disponibles.

- **Classificateur**

  Si tout s'est bien passé lors de l'analyse, le signe a été associé à une des entités présentes dans la zone d'attente et les valeurs de position et d'orientation fournies par le classificateur ont été récupérées et fournies à cette entité. Le module d'intégration supprime l'entité correspondante de la zone d'attente et ajoute un noeud décrivant cette entité, sa position et son orientation au graphe.

- **Verbe directionnel**

  Si tout s'est bien passé lors de l'analyse, le signe a été associé à plusieurs entités présentes dans le graphe. Selon le verbe traité, il va falloir ajouter une entité dans le graphe ou mettre à jour la position d'une entité dans le graphe, en utilisant les valeurs de positions initiale et finale du mouvement.

  Certaines entités voient leurs attributs évoluer au sein de la situation. Par exemple dans la phrase LSF signifiant "*Le garçon à ma droite me donne un gâteau*", l'entité *gâteau* se déplace du garçon vers le signeur : son emplacement varie. Pour chaque entité, il faut garder une représentation initiale et une représentation finale au sein de la situation, en supposant pour simplifier qu'une seule modification peut se produire. Ainsi, les entités sont représentées par un couple d'états (initial et final). Dans le cas du gâteau, l'état initial comporte les coordonnées du garçon et l'état final, les coordonnées du signeur.

La Figure 4.6 illustre la représentation de l'entité "gâteau" au sein du graphe.

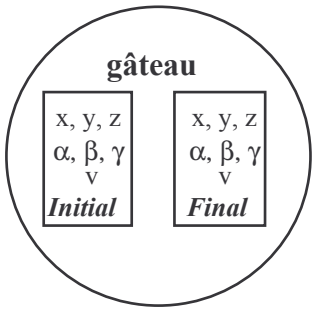


Figure 4.6 : Représentation de l'entité "gâteau".

- **Déictique**

  Ce type de signe ne provoque pas de modification dans le graphe. Il a pour fonction de désigner une des entités présentes dans la scène. Pour visualiser la mise en adéquation d'une entité avec un signe déictique, nous utilisons un attribut de mise en valeur (l'attribut *v* dans la Figure 4.6). Cet attribut peut par exemple servir à

augmenter la luminosité ou la taille d'une entité dans la scène virtuelle (décrite au Paragraphe 4.2.4), afin de visualiser si l'interprétation du signe déictique s'est effectuée correctement.

4.2.3. LES BASES DE DONNEES

4.2.3.1. **Base de données des Entités** (BDE)

Afin que les différents contrôles et mises à jour puissent être effectués, durant l'analyse ou l'intégration, il est nécessaire de disposer au niveau des entités d'informations sur le modèle de classificateur (plat, rond, personne) ainsi que sur leur statut (animé, inanimé). Une Base de Données des Entités (BDE) contient ces informations pour toutes les entités constituant le vocabulaire. De plus, la forme de ces entités est décrite au sein de cette base de donnée.

4.2.3.2. **Base de données des Variables** (BDV)

Par ailleurs, il est nécessaire de disposer d'une liste contenant l'ensemble des symboles correspondant aux signes variables. Pour chacun de ces symboles, les fonctions syntaxiques possibles sont indiquées. Dans le cas d'un classificateur, la "forme" des entités compatibles est spécifiée (plat, rond, personne…). Pour les verbes directionnels, le statut des entités compatibles est spécifié (par exemple un agent animé et un patient inanimé).

4.2.3.3. **Base de données des règles** (BDR)

Cette base de données contient toutes les règles utilisées par le processus d'Analyse-Contrôle et qui ont été décrites précédemment.

4.2.4. LA SCENE VIRTUELLE

La scène virtuelle représente la sortie du système. Il s'agit d'une image 3D représentant la scène de narration. A chaque étape de l'analyse, un fichier VRML (Virtual Reality Modeling Language) est créé à partir du graphe et de la zone d'attente. Il est visualisé à l'écran de l'ordinateur à l'aide d'un interpréteur de fichier VRML.

4.3. FONCTIONNEMENT : UN EXEMPLE

Pour illustrer le fonctionnement du système de compréhension, l'exemple de phrase en LSF présentée au chapitre précédent est repris. Il inclut deux signes standard, un classificateur et un verbe directionnel (Figure 4.7).

Figure 4.7 : Le signe "*garçon*" , un classificateur "index vertical" , le signe "*gâteau*" et le verbe directionnel "*donner*"

- Le premier signe "garçon" est un signe standard.
- Le second signe est un classificateur. Il est utilisé pour indiquer l'emplacement du garçon par rapport à l'emplacement du signeur dans la scène de narration. Les personnes debout sont en général représentées par ce type de classificateur.

  A cette étape de la phrase, on sait qu'il y a un garçon à la droite du signeur.
- Le troisième signe "gâteau" est un signe standard.
- Le quatrième signe "donner" est un verbe directionnel. Ici, un classificateur n'a pas été inclus. Il aurait été possible d'inclure un classificateur de type objet rond, reprenant ainsi l'entité "gâteau".

  Le mouvement part de la droite et va vers le signeur, ce qui signifie que le garçon donne quelque chose au signeur. Comme un gâteau a été signé juste avant, on peut en déduire que c'est un gâteau que le garçon donne au signeur.

Ainsi, la signification complète de la phrase en Français est : "*Le garçon qui est sur ma droite me donne un gâteau*".

Nous indiquons maintenant comment une telle phrase est traitée par le système de compréhension. Pour chaque étape, les opérations réalisées et les résultats obtenus sont indiqués.

Les figures représentent la scène virtuelle. Lorsqu'une entité est stockée dans le graphe, elle se situe dans la zone indiquée par le rectangle blanc. Lorsqu'elle est en attente, elle est placée dans le coin en haut à gauche de l'image.

ÉTAT INITIAL

Notons que le signeur est toujours présent dans la scène, dès l'état initial, puisque la scène de narration se construit en fonction de son emplacement.

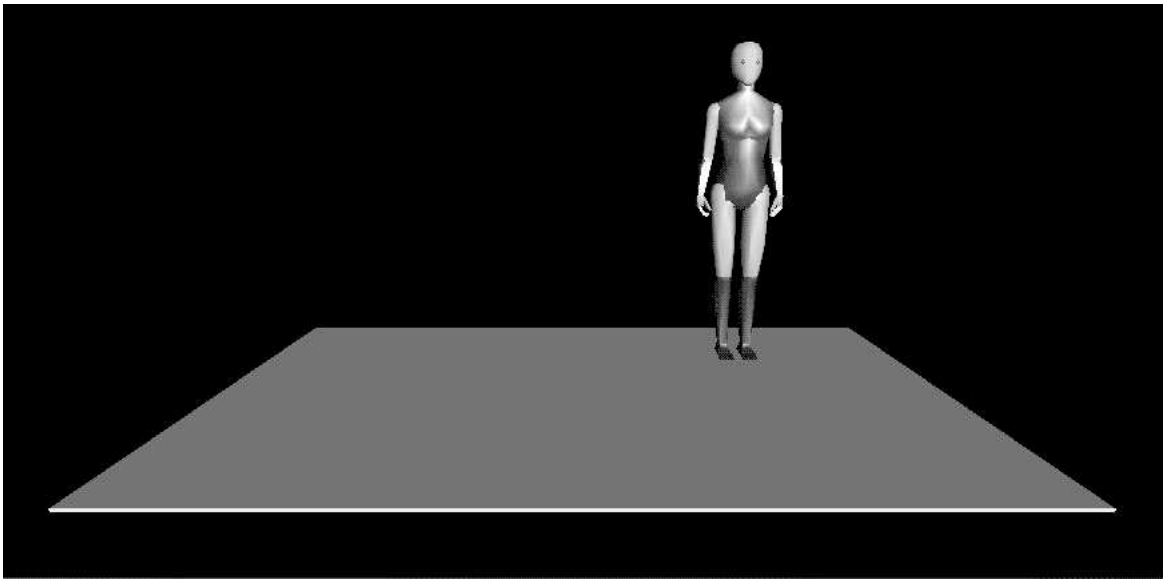


Figure 4.8 : État initial

PREMIER SIGNE

Reconnaissance -> "garçon"

Compréhension

• *Analyseur*

- Choix -> signe standard
- Analyse-Contrôle :

  Mise en attente de l'entité "garçon"
- Calcul-Contrôle : rien

• *Intégrateur* : rien

Scène virtuelle

Apparition de l'entité "garçon" en zone d'attente

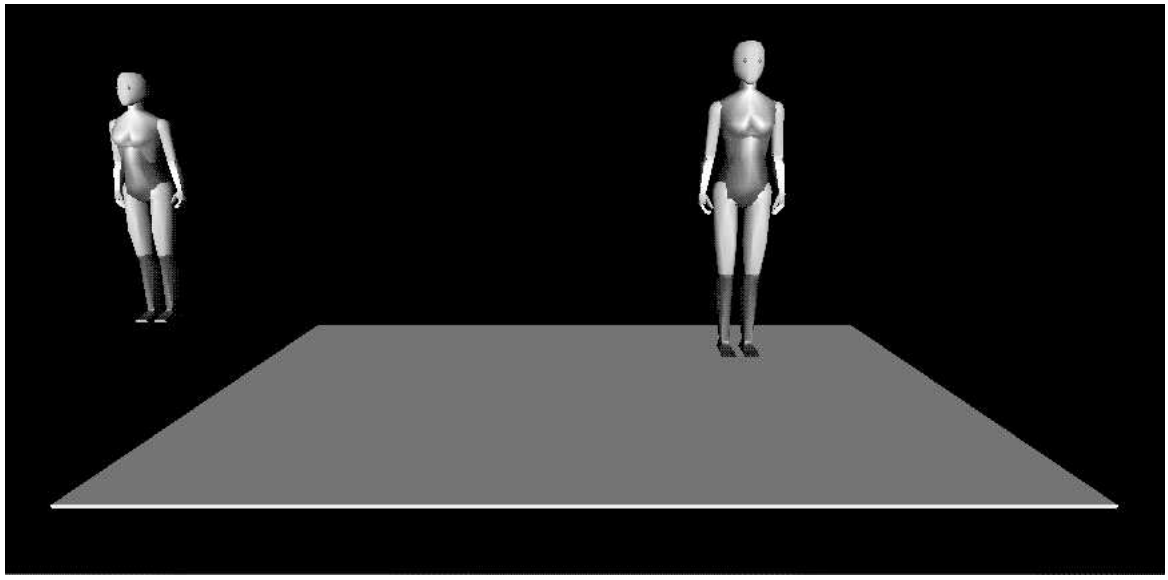
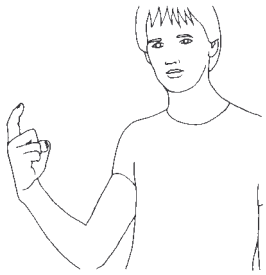


Figure 4.9 : Après l'interprétation du premier signe.

DEUXIEME SIGNE



Reconnaissance -> localisation

Compréhension

• *Analyseur*

- Choix -> signe variable
- Analyse-Contrôle :

  Recherche dans la BDV -> classificateur
  Recherche des l'entités -> garçon
  Contrôle de concordance classificateur/entité -> OK

- Calcul- Contrôle :

  Récupération des valeurs x, y, z et α, β, γ

• *Intégrateur*

Passage de l'entité "garçon" dans le graphe

Scène virtuelle

Intégration de l'entité "garçon" dans la scène de narration

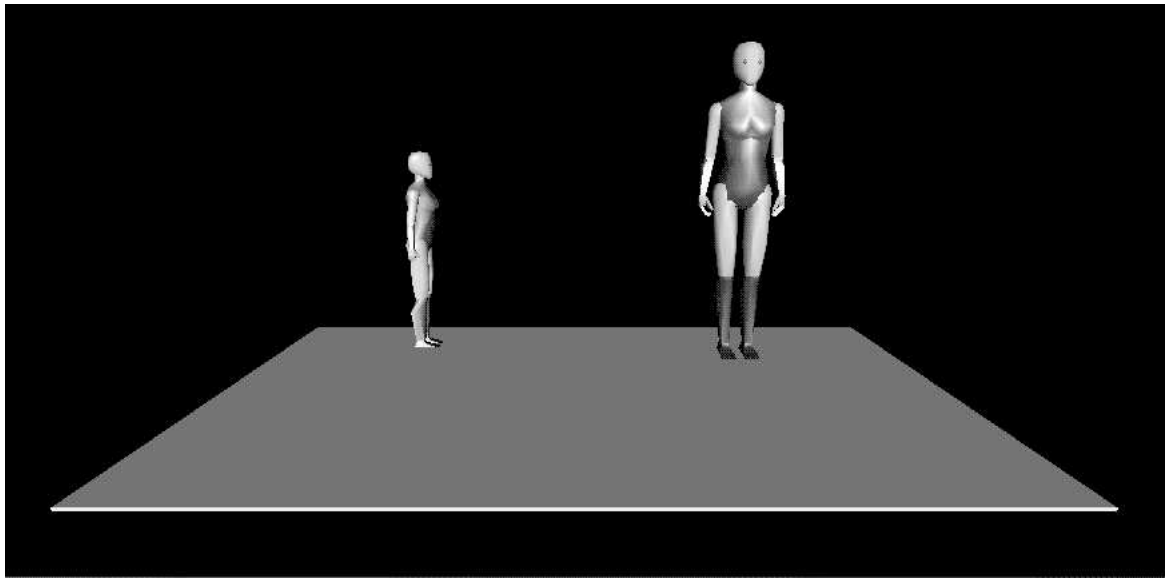


Figure 4.10 : Après l'interprétation du deuxième signe.

TROISIEME SIGNE

Reconnaissance -> "gâteau"

Compréhension

• *Analyseur*

- Choix -> signe standard
- Analyse-Contrôle :

  Mise en attente de l'entité "gâteau"
- Calcul-Contrôle : rien

• *Intégrateur* : rien

Scène virtuelle

Apparition de l'entité "gâteau" en zone d'attente

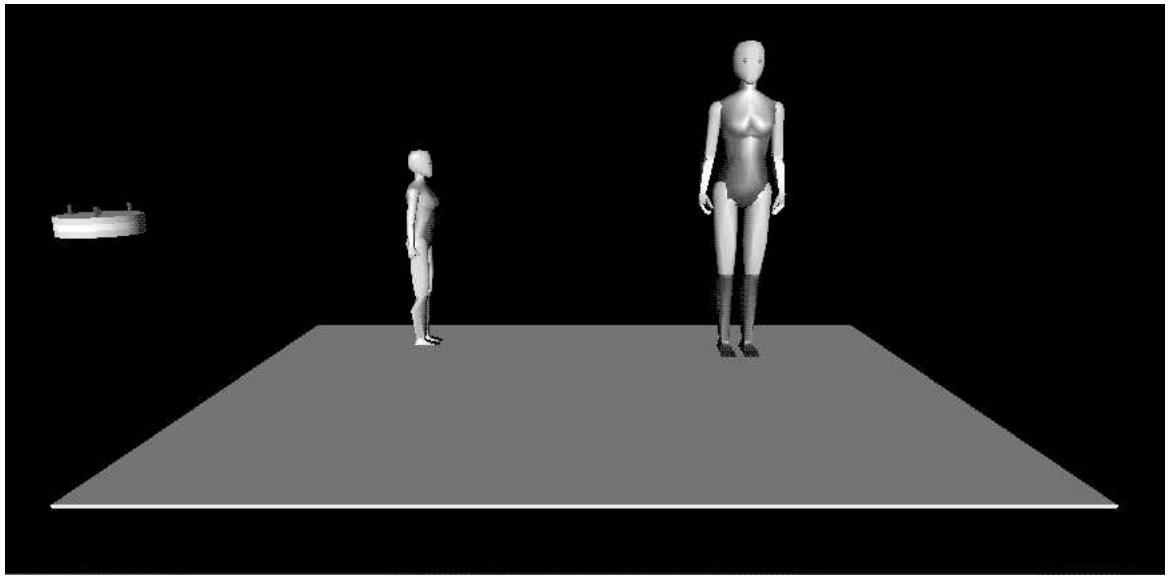


Figure 4.11 : Après l'interprétation du troisième signe.

QUATRIEME SIGNE

Reconnaissance -> "donner"

Compréhension

• *Analyseur*

- Choix -> signe variable
- Analyse-Contrôle :

  Recherche dans la BDV -> verbe directionnel
  Recherche des entités -> signeur, garçon, gâteau
  Contrôle des entités -> 2 animées, 1 inanimée : OK

- Calcul-Contrôle :

  Récupération des valeurs x_i, y_i, z_i et x_f, y_f, z_f
  Contrôle positions -> agent="garçon", patient="signeur"

• *Intégrateur*

Ajout de l'entité "gâteau" dans le graphe

Scène virtuelle

Intégration de l'entité "gâteau" dans la scène de narration
- d'abord à la position de l'agent (garçon) → Figure 4.12
- puis à la position du patient (signeur) → Figure 4.13

Ces deux étapes ont pout but de visualiser le résultat de l'action "donner" :

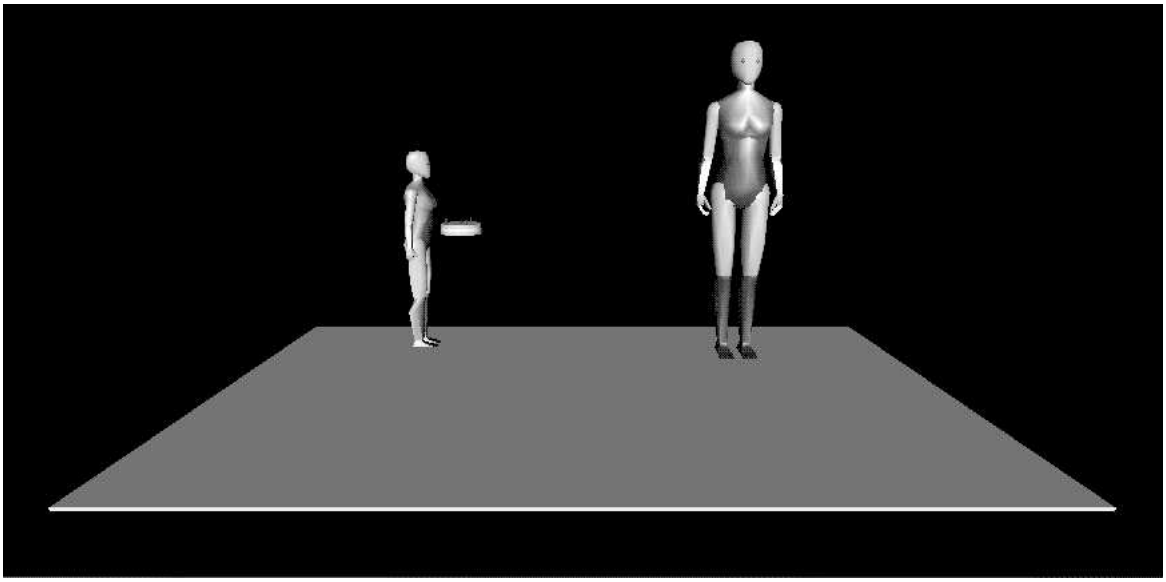


Figure 4.12 : Après l'interprétation du quatrième signe - 1ère étape.

Comme l'a illustré cet exemple, une image 3D est décrite à partir du graphe et de la zone d'attente et est affichée dans une fenêtre. Un module de traduction en français a été ajouté au prototype. Le résultat de la traduction apparait dans une fenêtre. Les deux fenêtres sont mises à jour à chaque étape du processus de compréhension. La Figure 4.13 montre le résultat final, à l'écran, de l'analyse de la phrase qui a été donnée en exemple. Par rapport à l'étape précédente, le gâteau est déplacé à la position du patient, c'est-à-dire le signeur.

Figure 4.13 : Après l'interprétation du quatrième signe - 2ème étape.

4.4. EVALUATION

Le prototype réalisé a été évalué à l'aide du corpus utilisé pour le module de reconnaissance. L'évaluation a consisté à vérifier pour chaque phrase du corpus si les entités sont correctement placées dans la scène et si les erreurs prévues sont détectées.

- Les phrases pour lesquelles aucune erreur n'a été produite durant le processus de reconnaissance ont été correctement analysées. Les entités sont placées au bon endroit dans la scène de narration et la traduction produite est correcte.
- Les erreurs de type insertion sont détectées :

  Par exemple, pour une des phrases [**garçon**] [**localisation**] [**personne**] [**aller vers**], une insertion du symbole garçon s'est produite (séquence à analyser : *garçon garçon localisation personne aller-vers*). Lors du processus de compréhension, le module d'analyse syntaxique détecte, pour le signe "localisation", que deux entités compatibles se trouvent dans la zone d'attente, ce qui est une erreur. Le processus d'analyse est stoppé à ce niveau.

  Pour quatre des phrases [**garçon**] [**localisation**] [**personne**] [**aller vers**], la séquence à analyser était *garçon localisation personne aller-vers localisation aller-vers*.

  Ce type d'erreur est détectée lors de l'analyse du deuxième classificateur localisation, pour lequel aucune entité ne peut être associée puisque la zone d'attente est vide. Le processus d'analyse est alors stoppé.
- Les erreurs de type suppression sont détectées :

  La seule phrase concernée est [**personne**] [**personne**] [**gâteau**] [**donner**], pour laquelle la séquence produite par le système de reconnaissance est *personne gâteau donner*. Ici, le graphe comporte deux entités compatibles, une personne et le signeur. L'erreur est détectée par comparaison des emplacements de ces entités et des valeurs relatives au verbe donner. Ici l'erreur est détectée à la fin de la phrase et le processus d'analyse est stoppé.

Sur le corpus utilisé, le prototype fonctionne parfaitement. Il traite avec succès les phrases correctement reconnues et est capable de détecter toutes les erreurs produites par le système de reconnaissance.

4.5. APPLICATIONS

Les applications possibles et les perspectives offertes par le système de reconnaissance et de compréhension ARGo sont de plusieurs types :

- L'application la plus directe concerne la modélisation de certains aspects du fonctionnement de la langue des signes. Les perspectives envisagées sont des outils dédiés à la LSF.
- Comme cela a été indiqué dans le Chapitre 1, les points communs entre les gestes co-verbaux et les gestes de la LSF sont multiples. Certaines parties du système ARGo pourraient être utilisées au sein d'applications, multimodales ou pas, utilisant la modalité gestuelle.
- ARGo peut être utilisé comme générateur de scène spatio-temporelle dans le cadre du système MoHA (Modèle Hybride d'Apprentissage) [Forest F. et Grau B. 1992], permettant de modéliser l'acquisition de connaissances sémantiques et pragmatiques par un système à partir des expériences perçues par ce système.

APPLICATIONS LIEES A LA LANGUE DES SIGNES

Si l'ordinateur possède un système de capture du geste, certaines applications dédiées à la langue des signes, s'appuyant sur le système de reconnaissance et de compréhension ARGo, peuvent être envisagées :

- Dictionnaire de signes avec entrée gestuelle.

  Actuellement, les dictionnaire de langue des signes proposent des accès aux représentations graphiques des signes par l'intermédiaire d'un index des traductions françaises par ordre alphabétique ou d'un index des signes organisé selon leurs configurations. De nouveaux outils informatiques proposent des séquences vidéo représentant les signes ainsi que des définitions, mais avec un index classique des traductions dans la langue orale correspondante. Il serait plus facile pour l'utilisateur d'accéder aux signes par l'intermédiaire d'une entrée gestuelle.

Le problème est que les systèmes de capture de gestes ne sont pas répandus à ce jour. En attendant qu'ils deviennent plus fiables et moins onéreux, il est possible de développer certaines applications prenant appui sur la partie compréhension d'ARGo. On peut envisager les applications suivantes :

- Génération de phrases de la LSF.

  A partir de la modélisation de la scène de narration d'ARGo, associée à un système de génération de gestes tel que celui développé au LIMSI [Lebourque T. et Gibet S. 1994].

- Tutoriaux d'aide à l'apprentissage de la syntaxe de la LSF à l'aide du français, ou dans l'autre sens, l'apprentissage de la syntaxe du français à l'aide de la LSF.

  Le système de compréhension présenté dans ce chapitre pourrait être un premier pas vers un système plus complet qui permettrait de mettre en oeuvre de tels tutoriaux. Cet objectif à plus long terme nécessite de développer un outil permettant de générer une phrase en français à partir du graphe d'ARGo et réciproquement. Cela a été réalisé dans le prototype dans le sens LSF $\rightarrow$ français, mais il faut l'étendre pour un plus grand vocabulaire.

INTERFACES GESTUELLES

Les gestes co-verbaux et les gestes de la LSF possèdent des points communs. Tout d'abord, pour les deux types de gestes, les paramètres permettent de transmettre des informations de type différent simultanément. La configuration permet de décrire un trait saillant de la forme d'une entité. Le mouvement permet de représenter une action ou de décrire le "profil" d'une forme. Le paramètre d'orientation indique l'orientation d'une entité dans la scène de narration. Enfin, l'emplacement permet de décrire les relations spatiales entre les entités.

Le module d'ARGo dédié à la reconnaissance peut être utilisé pour la reconnaissance de gestes de commande, par l'intermédiaire du module dédié aux signes standard, tandis que les gestes co-verbaux pourront être traités par le sous-module dédié aux signes variables.

Les applications correspondantes sont celles pour lesquelles des informations spatiales doivent être communiquées, par exemple des applications de description d'itinéraire ou de description d'une scène virtuelle (description d'arrangement mobilier, description d'objets).

CONSTRUCTION DE SITUATIONS POUR MOHA

Description de MoHA

MoHA (Modèle Hybride d'Apprentissage), proposé par F. Forest et B. Grau, permet de modéliser l'acquisition de connaissances sémantiques et pragmatiques à partir des expériences perçues par le système [Bordeaux F., Forest F. et al. 1992]. L'hypothèse de départ, tirée des travaux de Vygotsky sur les différentes étapes de la construction de concepts chez l'enfant, est que les relations de sens sont spécifiques à chaque individu et se construisent à partir de l'expérience du monde que cet individu a accumulée [Forest F. et Siksou M. 1994].

Ce modèle combine une approche numérique et une approche symbolique. L'approche numérique consiste à acquérir des expériences "vécues", nommées situations. Ces expériences peuvent provenir de différentes modalités, telles que la parole ou le geste. L'acquisition de ces situations débouche sur la formation de concepts. L'approche symbolique permet de construire des schémas représentant des informations de type pragmatique dans lesquels interviennent les concepts. L'apprentissage incrémental de ces schémas s'effectue à l'aide de traitements symboliques tels que l'analogie, la généralisation et la spécification [Ferret O. et Grau B. 1996].

Ce modèle comporte quatre niveaux, représentant les différents degrés d'intégration de l'expérience acquise par la machine. Les éléments des différents niveaux interagissent entre eux de manière à rendre le système dynamique. Les nouvelles perceptions peuvent participer au renforcement de concepts et de schémas existants, ceux-ci pouvant inversement servir à interpréter de nouvelles perceptions.

- Le premier niveau est celui qui représente l'acquisition de l'expérience à partir de perceptions (ou situations).
- Le deuxième niveau associe des "étiquettes linguistiques" aux entités participant aux situations, par exemple des chaînes phoniques dans le cas du langage oral.
- Le troisième niveau est un réseau sémantique classique au sein duquel les concepts sont reliés aux étiquettes linguistiques du deuxième niveau.
- Le quatrième niveau est composé d'un graphe de schémas représentant des situations prototypiques qui sont des généralisations d'événements particuliers perçus au niveau de l'expérience.

Chacun de ces niveaux représente un type différent d'abstraction. Le tout premier niveau ne procède à aucune abstraction, puisqu'il s'agit simplement de stocker des situations perçues. Ces situations sont représentées sous une forme proche de celle du graphe du système ARGo. C'est à ce niveau que notre système de compréhension peut être utilisé pour la construction du premier niveau de MoHA.

Utilisation du graphe d'ARGo pour MoHA[1]

Le niveau "perceptuel" de MoHA est constitué d'un graphe étroitement relié au deuxième niveau, celui des étiquettes linguistiques. Chaque noeud du graphe du premier niveau contient une situation, correspondant à une scène de narration telle que celle employée en LSF. Chaque entité présente dans la scène est reliée à un noeud du graphe de deuxième niveau, correspondant à une étiquette linguistique. Par exemple, la Figure 4.14 illustre la représentation aux niveaux 1 et 2 de MoHA de la situation suivante : l'agent donne un objet au patient. Les étiquettes linguistiques dans cet exemple sont des chaînes phoniques reliées aux entités "Je", "te" et "gâteau". La représentation spatio-temporelle permet d'indiquer les transformations dans la situation. Pour cela, chaque situation est représentée par une forme dans un espace multidimensionnel (espace 3D et temps) et chaque entité par un segment dans cet espace. Une même étiquette linguistique peut être reliée à plusieurs expériences.

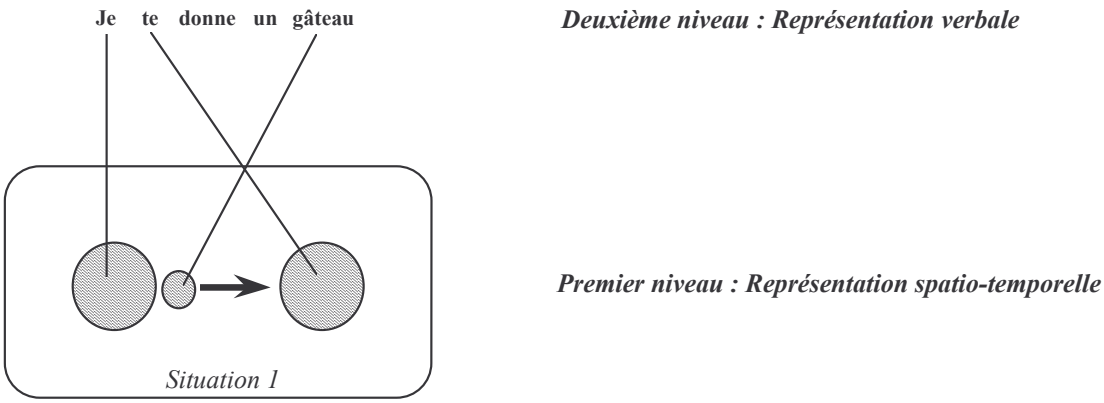


Figure 4.14 : Représentation d'une situation dans les graphes du 1er et 2ème niveau.

[1] Ce travail fait l'objet d'un projet de recherche, soutenu par une action incitative du LIMSI menée en collaboration avec Françoise Forest durant l'année 1995-1996, intitulée "*L'expression gestuelle vue comme une aide à la représentation informatique des connaissances*".

Pour construire son graphe de premier niveau, MoHA peut utiliser les graphes construits par ARGo. Ainsi, si dans le corpus mis en place pour ARGo il est possible de signer une scène telle que celle décrite dans l'exemple précédent, alors le graphe qui sera construit en fin du processus de compréhension pourra être ajouté au graphe de premier niveau de MoHA.

L'intérêt d'utiliser le graphe d'ARGo pour construire le premier niveau de MoHA est que les situations sont décrites directement sous forme de scènes dans l'espace, correspondant aux scènes perçues. Si la source d'acquisition des situations avait été limitée à des messages vocaux, il aurait fallu reconstruire la scène à partir d'une description verbale, ce qui est un sujet de recherche en soi.

Utilisation de MoHA pour ARGo

Une perspective à plus long terme est d'utiliser le système MoHA, quand il sera terminé, pour doter le système ARGo d'une approche descendante. En effet, les primitives perceptuelles sont reliées entre elles par des liens de proximité fonction de leur proximité perceptuelle dans les situations. En sortie du système de reconnaissance, dans le cas d'un signe standard représentant une entité, on pourrait explorer les primitives perceptuelles proches du noeud de MoHA représentant cette entité. On connaîtrait alors les entités ou les événements le plus souvent rencontrés. Cela permettrait d'optimiser la procédure de choix en sortie du système de reconnaissance.

4.6. CONCLUSION

Dans ce chapitre, nous avons proposé un système de compréhension de phrases de la LSF, connecté au système de reconnaissance décrit dans le Chapitre 3.

Il permet de compléter les informations fournies par le module de reconnaissance. Pour cela, il analyse la fonction syntaxique des signes de type variable afin de récupérer les valeurs d'emplacement, d'orientation ou de direction du mouvement. A l'aide de ces valeurs, l'interprétation de ces signes est complétée.

Le processus de compréhension repose sur la définition de règles spatio-temporelles relatives aux classificateurs, aux verbes directionnels et aux déictiques, ainsi que sur la modélisation de la scène de narration qui permet une prise en compte du contexte pour l'interprétation des signes variables. A chaque étape de l'interprétation, des contrôles sont effectués pour éviter de propager des erreurs d'interprétation.

L'implémentation d'un prototype a été réalisée et testée à l'aide du corpus utilisé pour l'évaluation du système de reconnaissance. Ce prototype inclus un module de traduction des phrases de LSF en français. Il est capable d'interpréter et de traduire les phrases qui ont été correctement reconnues par le système de reconnaissance et de détecter les erreurs lorsqu'elles se produisent.

CONCLUSION ET PERSPECTIVES

Dans cette thèse, nous avons présenté notre contribution aux recherches sur l'étude du mode gestuel et son utilisation en communication homme-machine.

Rares sont les études portant sur le geste humain en vue de proposer une modélisation informatique. C'est pourquoi une partie importante du travail exposé dans cette thèse a consisté à étudier ses propriétés. Une première étude comparative a permis de dégager des propriétés concernant la structure des gestes de la main communes aux gestes co-verbaux et aux signes de la langue des signes.

Les gestes de la main, possèdent quatre composantes (nommés paramètres en LSF). Ces quatre paramètres sont la configuration, le mouvement, l'orientation et l'emplacement de la main. Chacun d'eux est chargé de véhiculer une information d'un type différent. Ces quatre types d'information peuvent être émis simultanément. De plus, le locuteur ou le signeur utilise une scène de narration située devant lui et c'est en plaçant des entités au sein de cette scène qu'il construit l'image contenant l'information qu'il veut communiquer. C'est le moyen utilisé pour représenter le contexte, afin de résoudre l'ambiguïté le message.

La similitude entre les gestes co-verbaux et les gestes de la LSF nous a poussé à mener une étude plus approfondie de ces derniers, à partir d'un corpus "papier" de 1257 signes. Nous avons constaté que pour la reconnaissance, deux catégories de signes doivent être distinguées : ceux pour lesquels les quatre paramètres sont invariables quelque soit le contexte (les signes standard) et ceux pour lesquels au moins un des paramètres est variable en fonction du contexte (les signes variables). Cette dernière catégorie inclut les classificateurs et les verbes directionnels. Pour la compréhension, deux catégories de paramètres doivent être différenciés selon qu'ils possèdent une valeur sémantique ou pas. Les paramètres Configuration et Mouvement possèdent toujours une valeur sémantique, ce qui n'est pas toujours le cas des paramètres Orientation et Emplacement. En revanche, lorsque ces deux derniers paramètres sont porteurs d'information, leurs valeurs numériques précises sont nécessaires pour compléter l'interprétation.

Fondé sur ces remarques, nous proposons un système de reconnaissance de phrases de la LSF composées de signes standard et variables. La technique de reconnaissance utilisée est basée sur les modèles de Markov cachés qui ont permis l'obtention de taux de reconnaissance très encourageants sur les deux catégories de signes (96% pour les signes standard et 92% pour les signes variables). L'évaluation porte sur un corpus constitué de deux ensembles de

44 phrases différentes composées de quatre signes. Les phrases gestuelles ont été réalisées par une seule personne. Le premier ensemble a été utilisé pour l'apprentissage et le deuxième, pour la reconnaissance.

Le système de compréhension proposé est également capable de traiter les signes standard et variables. Il repose sur la définition de règles spatio-temporelles relatives aux classificateurs, aux verbes directionnels et aux déictiques, ainsi que sur la modélisation de la scène de narration qui permet une prise en compte du contexte pour l'interprétation des signes variables. La réalisation d'un prototype auquel a été adjoint un module de traduction a permis d'évaluer le système complet (reconnaissance et compréhension), sur le corpus utilisé pour le module de reconnaissance. Les traduction fournies sont correctes et les erreurs produites par le module de reconnaissance sont détectées.

Malgré ses limitations dues en partie au système de capture de geste (le DataGlove) et en partie à l'outil utilisé pour construire le système de reconnaissance (la version actuelle des programmes de mise en oeuvre des HMM), le système ARGo est une avancée importante dans le domaine de la reconnaissance et compréhension de phrases gestuelles, car il permet de traiter à la fois des signes standard, des classificateurs, des verbes directionnels et des déictiques. Le graphe de représentation de la scène de narration, visualisé par l'intermédiaire de la scène virtuelle, est un premier pas vers un système de représentation d'informations transmises par l'intermédiaire du canal gestuel, et dont les applications vont du domaine général de l'interaction gestuelle au domaine plus spécifique des applications dédiées à la langue des signes.

Les évolutions possibles du système sont multiples. Plusieurs sont envisagées à court terme :

- La première amélioration à apporter au système concerne l'outil de reconnaissance. Cet outil doit permettre d'une part de donner un poids identique aux paramètres co-occurrents et d'autre part de disposer des scores de reconnaissance par signes. Cela favorisera l'augmentation des performances du système et autorisera l'implémentation du processus de choix du module de compréhension.
- La deuxième évolution concerne la capture du mouvement de la main gauche. Pour les signes utilisant les deux mains, nous avons présenté dans le Chapitre 2 les rapports entre les mains, aussi bien du point de vue configuration que du point de vue mouvement. Ces rapports assez simples laissent supposer qu'il n'est pas nécessaire de disposer d'un gant pour capter les mouvements de la main dominée.

Une caméra ou un capteur de type Polhemus devrait suffire. La capture de la main dominée permettra d'étendre considérablement le vocabulaire et le type d'informations transmises. Par exemple, il sera possible d'exprimer les relations spatiales entre les entités, ce qui n'est pas le cas à l'heure actuelle.

- La troisième évolution est l'ajout d'un module d'animation 3D pour visualiser les composantes dynamiques présentes dans la scène de narration. La visualisation dont nous disposons actuellement est statique. Or une scène de narration est par essence dynamique. Il serait plus informatif de voir se déplacer les entités dynamiques du discours.
- Des systèmes simples permettent de capter des pressions. Comme cela a été indiqué dans le Chapitre 2, les contacts sont des composantes importantes qui sont présentes dans les configurations dynamiques et au niveau des interactions entre les mains et certaines parties du corps. Nous envisageons d'ajouter ce type de capteurs au système. Ils devraient permettre de pallier au moins en partie le manque de précision du gant et d'augmenter le nombre de configurations différenciables et donc la taille du vocabulaire.

D'autres évolutions sont envisagées à plus long terme :

- Lorsque le système de capture sera suffisamment précis et fiable pour capter une grande quantité de signes différents, il sera possible de mettre en place une collaboration avec des personnes dont la langue maternelle est la LSF afin de construire des corpus plus importants et représentatifs de cette langue.
- L'ajout d'un module de description de la scène de narration plus général que celui qui a été développé dans le prototype permettra de traduire une phrase de LSF en français. Ce travail devra être réalisé en collaboration avec des spécialistes du traitement automatique du langage naturel. Ce développement sera une nouvelle étape vers la mise en oeuvre de systèmes permettant de construire des didacticiels permettant l'apprentissage d'une des langues à l'aide de l'autre.
- En parallèle aux études relatives à la LSF, les principes à partir desquels le système ARGo a été construit pourront être utilisés pour créer des systèmes de reconnaissance et de compréhension de gestes intégrés dans des applications multimodales. En effet, les signes de type standard peuvent être rapprochés des gestes emblématiques, tandis que les signes variables peuvent être rapprochés des gestes illustrateurs. Ces deux types de gestes sont utilisés dans le contexte de la communication multimodale.

BIBLIOGRAPHIE

[Baecker R. M. et Buxton W. A. S. 1987]
Baecker R. M. et Buxton W. A. S. (1987). The Haptic Channel. Readings in Human-Computer Interaction. A Multidisciplinary Approach, pp.357-365.

[Baudel T. et Braffort A. 1992]
Baudel T. et Braffort A. (1992). *Reconnaissance de gestes de la main pour l'aide à la présentation assistée par ordinateur.* Document vidéo LIMSI k92-04.

[Baudel T. et Braffort A. 1993]
Baudel T. et Braffort A. (1993). *Reconnaissance des gestes de la main en environnement réel*. Actes de Informatique'93. L'interface des mondes réels et virtuels, EC2, pp.207-216, Montpellier.

[Bellalem N. 1995]
Bellalem N. (1995). *Etude du mode de désignation dans un dialogue homme-machine finalisé à forte composante langagière : analyse structurelle et implémentation*. Thèse de doctorat d'informatique, Université Nancy 1.

[Bellik Y. 1991]
Bellik Y. (1991). *Interface de dialogue multimodal*. Mémoire de DEA, Orsay.

[Bellik Y., Pican N. et al. 1994]
Bellik Y., Pican N. et Burger D. (1994). *Méditor, un prototype d'interface multimodale pour la manipulation de textes braille enrichis*. Interfaces multimodales pour handicapés visuels, Numéro spécial de la revue "Comme les autres", pp.47-59, Novembre 1994.

[Bellugi U. et Klima E. 1979]
Bellugi U. et Klima E. (1979). *The Signs of Language*. (Harvard University Press), Cambridge.

[Boehm K., Broll W. et al. 1994]
Boehm K., Broll W. et Sokolewicz M. (1994). *Dynamic Gesture Recognition using Neural Networks; A Fundament for Advanced Interaction Construction*. Actes de SPIE Conference Electronic Imaging Science & Technology, San Jose, California, USA, Février 1994.

[Bolt R. A. 1980]

Bolt R. A. (1980). *"Put that there": Voice and Gesture at the Graphics Interface*. Computer Graphics, vol.14, pp.262-270.

[Bolt R. A. 1987]

Bolt R. A. (1987). *Conversing with Computers*. Readings in Human-Computer Interaction - A multidisciplinary approach, (Morgan Kaufmann), pp.694-702.

[Bordeaux F., Forest F. et al. 1992]

Bordeaux F., Forest F. et Grau B. (1992). *MoHA, an hybrid learning model: a model based on the perception of the environment by an individual*. Actes de IPMU'92, Majorque (Espagne).

[Bordegoni M. et Hemmje M. 1993]

Bordegoni M. et Hemmje M. (1993). *An Interaction Model Based on Hand Gestures for 3D User Interfaces*. Actes de Workshop ERCIM on Multimodal Human-Computer Interaction, INRIA-Lorraine, Nancy.

[Bourdot P., Krus M. et al. 1995]

Bourdot P., Krus M. et Gherbi R. (1995). *MIX 3D: une plate-forme expérimentale pour des interfaces multimodales dédiées à la CAO*. Document vidéo, IHM'95.

[Braffort A. 1992]

Braffort A. (1992). *Définition d'un modèle d'interaction pour gant numérique*. Mémoire de DEA, Orsay.

[Braffort A. 1996a]

Braffort A. (1996a). *ARGo: An Architecture for Sign Language Recognition and Interpretation*. Actes de Gesture Workshop'96, York (GB), Mars 1996.

[Braffort A. 1996b]

Braffort A. (1996b). *A gesture recognition architecture for sign language*. Actes de ASSETS'96, ACM, Vancouver (Canada), Avril 1996.

[Braffort A., Baudel T. et al. 1992]

Braffort A., Baudel T. et Teil D. (1992). *Utilisation des gestes de la main pour l'interaction homme-machine*. Actes de IHM'92, GDR-PRC-CHM, pp.193-196, Paris.

[Braffort A., Collet C. et al. 1994a]

Braffort A., Collet C. et Teil D. (1994a). *Anthropomorphic Model for Hand Gesture Interface*. Actes de CHI'94, ACM, pp.259-260, Boston.

[Braffort A., Collet C. et al. 1994b]

Braffort A., Collet C. et Teil D. (1994b). *Hand configuration pre-processing tool for Sign Language Recognition*. Actes de RESNA'94, Nashville, Tennessee.

[Briffault X. 1992]

Briffault X. (1992). *Modélisation informatique de l'expression de la localisation en langage naturel*. Thèse de doctorat d'informatique, Orsay.

[Briffault X. et Braffort A. 1993a]

Briffault X. et Braffort A. (1993a). *Toward a Model of Cooperation Between Natural Language and Natural Gestures to Describe Spatial Knowledge*. Actes de PacLing'93, Vancouver (Canada).

[Briffault X. et Braffort A. 1993b]

Briffault X. et Braffort A. (1993b). *Space, Language and Gestures: A Model of Multimodal Expression of Space*. Actes de IASTED'93, pp.156-159, Annecy.

[Buxton W. 1987]

Buxton W. (1987). *There's More to Interaction than Meets the Eye: Some Issues in Manual Input*. Readings in Human-Computer Interaction. A Multidisciplinary Approach.

[Cadoz C. 1994]

Cadoz C. (1994). *Le geste canal de communication homme/machine - la communication "instrumentale"*. Techniques et Science Informatiques, vol.13, pp.31-61.

[Cagin J.-M. 1993]

Cagin J.-M. (1993). *Une étude sur la reconnaissance de formes dynamiques - Application à la langue des signes*. Projet de fin d'études de l'ENSTA, Paris.

[Calais-Germain B. 1989]

Calais-Germain B. (1989). *Anatomie pour le mouvement - Introduction à l'analyse des techniques corporelles*. ISBN : 2-9500608-0-3.

[Calbris G. 1985]

Calbris G. (1985). *Espace-temps : Expression gestuelle du temps*. Semiotica, vol.55, pp.43-73.

[Calbris G. 1993]

Calbris G. (1993). *Le geste co-verbal expression du temps.* Document vidéo LIMSI n°29.

[Calbris G. et Montredon J. 1986]

Calbris G. et Montredon J. (1986). *Des gestes et des mots pour le dire*. (Dic Mini-dictionnaires).

[Cassell J., Pelachaud C. et al. 1994]

Cassell J., Pelachaud C., Badler N., Steedman M., Achorn B., Becket T., Douville B., Prevost S. et Stone M. (1994). *Animated conversation: Rule-based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conversational Agents*. Actes de SIGGRAPH'94, Orlando, USA.

[Collet C. 1993]

Collet C. (1993). *Reconnaissance de gestes par réseaux neuromimétiques*. Mémoire de DEA, Orsay.

[Cuxac C. 1983]

Cuxac C. (1983). *Autour de la Langue des Signes*. vol.10, Journée d'études n°10, (Université Paris V, UER de linguistique générale et appliquée), Paris.

[Cuxac C. 1987]

Cuxac C. (1987). *La transitivité et ses corrélats*. Cycle de conférences, centre de linguistique, Travaux n°1.

[Cuxac C. 1993a]

Cuxac C. (1993a). *Iconicité des langues des signes*. Actes de 4eme école d'été de L'ARC - Communication et Multimodalité dans les Systèmes Naturels et Artificiels, GDR-PRC-CHM ARC, pp.203-216, Bonas (Gers).

[Cuxac C. 1993b]

Cuxac C. (1993b). *L'expression du temps dans la langue des signes.* Document vidéo LIMSI n°29.

[Da Silva Faria O. 1991]

Da Silva Faria O. (1991). *Reconnaissance gestuelle à partir d'un gant numérique*. Mémoire de DEA, Orsay.

[De Fornel M. 1993]

De Fornel M. (1993). *Sémantique et pragmatique du geste métaphorique*. Cahiers de linguistique française, vol.14, pp.247-253.

[Delannoy J. F. et Lula J. B. 1990]

Delannoy J. F. et Lula J. B. (1990). *Une voie prometteuse de la communication homme-machine : les interfaces graphiques*. vol.16, (Interactions homme-machine), Rouen.

[Dubois J., Giacomo M. et al. 1973]

Dubois J., Giacomo M., Guespin L., Marcellesi C., Marcellesi J.-B. et Mevel J.-P. (1973). *Dictionnaire de linguistique*. (Larousse), Paris.

[Duda R. O. et Hart P. E. 1973]

Duda R. O. et Hart P. E. (1973). *Pattern Classification and Scene Analysis*. (Wiley).

[Eglowstein H. 1990]

Eglowstein H. (1990). *Reach Out and Touch Your Data*. BYTE, Juillet 1990.

[Ekman P. et Friesen W. V. 1972]

Ekman P. et Friesen W. V. (1972). *Hand Movements*. The Journal of Communication, vol.22, pp.353-374.

[Ellis S. R. 1994]

Ellis S. R. (1994). *What are Virtual Environments ?* IEEE Computer Graphics & Applications, pp.17-22, Janvier 1994.

[Encarnação J., Göbel M. et al. 1994]

Encarnação J., Göbel M. et Rosenblum L. (1994). *European Activities in Virtual Reality*. IEEE Computer Graphics & Applications, pp.66-74, Janvier 1994.

[Erenshteyn R., Foulds R. et al. 1994]

Erenshteyn R., Foulds R. et Galuska S. (1994). *Is Designing a Neural Network Application an Art or a Science*. SIGCHI Bulletin, vol.26, pp.23-29.

[Faure C. et Julia L. 1992]

Faure C. et Julia L. (1992). *TAPAGE : une interface pour l'aide à l'édition de TAbleaux par la PArole et le GEste*. Actes de IHM'92 Quatrièmes journées sur l'ingénierie des interfaces homme-machine, GDR_PRC_CHM, pp.167-171, Paris.

[Fels S. S. 1994]

Fels S. S. (1994). *Glove-TalkII: Mapping Hand Gestures to Speech Using Neural Networks - An Approach to Building Adaptative Interfaces*. PhD, Toronto.

[Fels S. S. et Hinton G. E. 1990]

Fels S. S. et Hinton G. E. (1990). *Building Adaptative Interfaces with Neural Networks: The Glove-Talk Pilot Study*. Actes de Human-Computer Interaction - INTERACT'90, IFIP, pp.683-688.

[Fels S. S. et Hinton G. E. 1993]

FELS S. S. et HINTON G. E. (1993). *Glove-Talk: A Neural Network Interface Between a Data-Glove and a Speech Synthesizer*. IEEE Transactions on Neural Networks, vol.4, pp.2-8.

[Ferret O. et Grau B. 1996]

Ferret O. et Grau B. (1996). *Construire une mémoire épisodique à partir de textes : pourquoi et comment ?* Actes de RFIA'96, Rennes (France).

[Foley 1987]

Foley (1987). *Les communications entre l'Homme et l'Ordinateur*. Pour la science, Décembre 1987.

[Foley J., Van Dam A. et al. 1990]

Foley J., Van Dam A., Feiner S. K. et Hughes J. F. (1990). *Computer graphics Principles and practice*. (Addison Wesley).

[Forest F. et Grau B. 1992]

Forest F. et Grau B. (1992). *MoHA, un Modèle Hybride d'Apprentissage*. Actes de Journées de Rochebrune, AFCET, ARC, Janvier 1992.

[Forest F. et Siksou M. 1994]

Forest F. et Siksou M. (1994). *Développement de concepts et programmation du sens, Pensée et Langage chez Vygotsky*. Intellectica, vol.1, n°18, pp.213-236.

[Friedman L. A. 1975]

Friedman L. A. (1975). *Space, Time, and Person Reference in American Sign Language*. Language, vol.4.

[Gauvain J. L., Lamel L. et al. 1994]

Gauvain J. L., Lamel L., Adda G. et Adda-Decker M. (1994). *Speaker-independent continuous speech dictation*. Speech Communication, vol.15, n°1-2, pp.21-37.

[Geoffrois E. 1995]

Geoffrois E. (1995). *Extraction robuste de paramètres prosodiques pour la reconnaissance de la parole*. Thèse de doctorat d'informatique, Université d'Orsay.

[Gibet S. 1987]

Gibet S. (1987). *Codage, représentation et traitement du geste instrumental - Application à la synthèse de sons musiquaux par simulation de mécanismes instrumentaux*. Thèse de doctorat d'informatique, INPG Grenoble.

[Gibet S. 1992]

Gibet S. (1992). *Nonlinear Feedback Model of Sensori-Motor Systems*. Actes de International Conference on Automation, Robotics and Computer Vision (ICARV'92), Singapore.

[Gibet S. et Marteau P.-F. 1994]

Gibet S. et Marteau P.-F. (1994). *A Self-organized Model for the Control, Planning and Learning of Nonlinear Multidimentional Systems using a Sensory Feedback*. Applied Intelligence, n°4, pp.337-349.

[Gourley C. 1994]

Gourley C. (1994). *Neural Networks Utilizing Posture Input for Sign Language Recognition*. Rapport techique, Computer Vision & Robotics Research Laboratory - University of Tenessee - Knoxville, 28 Novembre 1994.

[Hand C., Sexton I. et al. 1994]

Hand C., Sexton I. et Mullan M. (1994). *A Linguistic Approach to the Recognition of Hand Gestures*. Actes de Designing Future Interaction, Ergonomics Society/IEE, University of Warwick, UK, Avril 1994.

[Harling P. A. 1993]

Harling P. A. (1993). *Gesture Input using Neural Networks*. BSc degree in Computer Science, Dept of Computer Science University of York.

[Jodouin J.-F. 1994]

Jodouin J.-F. (1994). *Les réseaux de neurones. Principes et définitions*. (Hermès).

[Jodouin J.-F. 1994]

Jodouin J.-F. (1994). *Les réseaux neuromimétiques. Modèles et applications*. (Hermès).

[Kadous W. 1995]

Kadous W. (1995). *GRASP: Recognition of Australian Sign Language using Instrumented Gloves*. Bachelor of Computer Engineering, University of New South Wales.

[Kahaner D. 1994]

Kahaner D. (1994). *Japanese Activities in Virtual Reality*. IEEE Computer Graphics & Applications, pp.75-78, Janvier 1994.

[Kendon A. 1980]

Kendon A. (1980). *Gesticulation and Speech: Two Aspects of the Process of Uterance*. The Relation between Verbal and Non-verbal Communication, (M.R. Key), Mouton, The Hauge, pp.207-227.

[Kramer J. et Leifer L. 1989]

Kramer J. et Leifer L. (1989). *The "Talking Glove": A speaking Aid for Nonvocal Deaf and Deaf-blind Individuals*. Actes de RESNA 12th, pp.471-472, New Orleans (Louisiana).

[Kurtenbach G. et Buxton B. 1991]

Kurtenbach G. et Buxton B. (1991). *GEdit : A Test Bed for Editing by Contiguous Gestures*. SIGCHI Bulletin, vol.23, pp.22-26.

[Lafouillade E. 1992]

Lafouillade E. (1992). *3D Hand Animation for Sign Languages*. Mémoire de fin d'études, IIE (Evry).

[Lane H., Boyes-Braem P. et al. 1976]

Lane H., Boyes-Braem P. et Bellugi U. (1976). *Preliminaries to a Distinctive Feature Analysis of Handshapes in American Sign Language*. Cognitive Psychology, vol.8, pp.263-289.

[Latta J. N. et Oberg D. J. 1994]

Latta J. N. et Oberg D. J. (1994). *A conceptual Virtual Reality Model*. IEEE Computer Graphics & Applications, pp.23-29, Janvier 1994.

[Lebourque T. et Gibet S. 1994]

Lebourque T. et Gibet S. (1994). *Un modèle de génération de gestes naturels*. Actes de IHM'94, Ganymède PRC-CHM USTL, Trigone, pp.107-112, Lille, 8 et 9 Décembre 1994.

[Lee J. 1994]

Lee J. (1994). *Notational Representation of Sign Language: A Structural Description of Hand Configuration*. Actes de ICCHP'94, Springer-Verlag, pp.38-45, Vienne.

[Lee J. et Kunii T. L. 1993]

Lee J. et Kunii T. L. (1993). *Computer Animated Visual Translation From Natural Language to Sign Language*. The journal of visualization and computer animation, vol.4, pp.63-78.

[Liang R.-H. et Ming O. 1995]

Liang R.-H. et Ming O. (1995). *A Real-time Continuous Alphabetic Sign Language to Speech Conversion VR System*. Actes de Eurographics'95.

[Liddell S. K. 1990]

Liddell S. K. (1990). *Structures for Representing Handshape and Local Movement at the Phonemic Level*. Theorical Issues in Sign Language Research, vol.1 : Linguistics, (The University of Chicago Press), Chicago & London.

[Loomis J., Poizner H. et al. 1983]

Loomis J., Poizner H., Bellugi U., Blakemore A. et Hollerbach J. (1983). *Computer Graphic Modeling of American Sign Language*. Computer Graphics, vol.17, pp.105-114.

[LRP 1993]

LRP (1993). *Rapport d'activités 1990-1993* Laboratoire de Robotique de Paris, Juin 1993.

[Mariani J. 1993]

Mariani J. (1993). *Automated Voice Dictation in French*. Speech Communication, vol.13, n°1-2, pp.171-185.

[Marr D. 1982]

Marr D. (1982). *Vision*. (Freedman and company).

[Martin-Dupont X. 1995]

Martin-Dupont X. (1995). *Les modalités d'évaluation objective dans le domaine de la communication non verbale*. Notes et documents, LIMSI, 95-08, Mars 1995.

[Mc Neill D. 1992]

Mc Neill D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. (The University Chocago Press), Chicago.

[Messing L. S., Erenshteyn R. et al. 1994]

Messing L. S., Erenshteyn R., Foulds R., Galuska S. et Stern G. (1994). *American Sign Language Computer Recognition: Its Present and its Promise*. Actes de ISAAC'94, pp.289-291, Maastricht, NL.

[Miclet L. 1984]

Miclet L. (1984). *Méthodes structurelles pour la reconnaissance de formes*. (Collection technique et scientifique des télécommunications), Paris.

[Moody B. 1983]

Moody B. (1983). *La langue des signes. Histoire et grammaire*. vol.1, Paris.

[Moody B. 1986]

Moody B. (1986). *La langue des signes. Dictionnaire bilingue élémentaire*. vol.2, Paris.

[Morrel-Samuels P. 1990]

Morrel-Samuels P. (1990). *Clarifying the Distinction between Lexical and Gestural Commands.* International Journal of Man-Machine Studies, vol.32, pp.581-590.

[Murakami K. et Taguchi H. 1991]

Murakami K. et Taguchi H. (1991). *Gesture Recognition using Recurrent Neural Networks*. Actes de CHI'91, ACM, pp.237-242, New Orleans (Louisiana).

[Nam Y. et Wohn K. 1995]

Nam Y. et Wohn K. (1995). *Recognition of Space-Time Hand-Gestures using Hidden Markov Model*. Actes de VRAIS'95.

[Newby G. B. 1993]

Newby G. B. (1993). *Gesture recognition using Statistical Similarity*. Actes de Virtual Reality abd Persons with Disabilities.

[Nogier J.-F. 1993]

Nogier J.-F. (1993). *Dialogue Homme-Machine Multimodal. Application aux systèmes de contrôle & Surveillance*. Actes de Informatique'93. L'interface des mondes réels et virtuels, EC2, pp.191-204, Montpellier.

[Perlmutter D. M. 1990]

Perlmutter D. M. (1990). *On the Segmental Representation of Transitional and Bidirectional Movements in ASL Phonology*. Theorical Issues in Sign Language Research, vol.Volume 1: Linguistics, (University of Chicago Press), Chicago & London.

[Poizner H., Klima E. S. et al. 1986]

Poizner H., Klima E. S., Bellugi U. et Livingston R. B. (1986). *Motion Analysis of Grammatical Processes in a Visual-Gestural Language*. Motion: Representation and Perception, ACM, North-Holland.

[Prillwitz S. et Leven R. 1989]

Prillwitz S. et Leven R. (1989). *HAMNOSYS Version 2.0*. (SIGNUM PRESS), Hamburg.

[Rabiner L. R. 1989]

Rabiner L. R. (1989). *A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*. IEEE, vol.77-2, pp.257-285, Février 1989.

[Revesz P. Z. et Raghava-Rao V. K. 1993]

Revesz P. Z. et Raghava-Rao V. K. (1993). *A Sign-to-speech Translation System Using Matcher Neural Networks*. Actes de Conference on Artificial Neural Networks in Engineering, St Louis (Missouri).

[Roskos E. M. et Zhuang J. 1990]

Roskos E. M. et Zhuang J. (1990). *Real Time Software for the VPL DataGlove*. Rapport technique, CAIP Computer Aids for Industrial Productivity, CAIP-TR-124.

[Rubine D. 1991a]

Rubine D. (1991a). *The automatic recognition of gestures*. PhD thesis, Carnegie Mellon University.

[Rubine D. 1991b]

Rubine D. (1991b). *Specifying Gestures by Example*. Computer Graphics, vol.25, pp.329-337.

[Sagawa H., Sakou H. et al. 1992]

Sagawa H., Sakou H. et Abe M. (1992). *Sign Language Translation System Using Continuous DP Matching*. Actes de MVA'92 - IAPR Workshop on Machine Vision Applications, pp.339-342, Tokyo.

[Sandler W. 1989]

Sandler W. (1989). *Phonological Representation of the Sign - Linearity and Nonlinearity in American Sign Language*. (Foris Publications), Dordrecht (Holland) & Providence (USA).

[Searles D., Smith J. et al. 1993]

Searles D., Smith J., Baratoff G. et Bohmueller B. (1993). *Recognition of Signals for Combat Formations and Battle Drills*. Rapport électronique JOVE (Journal of Virtual Recognition), Dept of Computer Science University of Maryland.

[Shneiderman B. 1987]

Shneiderman B. (1987). *Direct Manipulation : A Step Beyond Programming Languages*. R.M. Beacker & B. Buxton, (a multidisciplinary approach Readings in human-computer interaction), pp.461-467.

[Simon J. C. 1984]

Simon J. C. (1984). *La reconnaissance des formes par algorithmes*. J. Berstel, (Collection ERI (études et recherches en informatique)), Paris.

[Sparrel C. J. 1993]

Sparrel C. J. (1993). *Coverbal Iconic Gesture in Human-Computer Interaction*. Master of Science, MIT.

[Starner T. et Pentland A. 1995]

Starner T. et Pentland A. (1995). *Visual Recognition of American Sign Language Using Hidden Markov Models*. Actes de International Workshop on Automatic Face and Gesture Recognition, Zurich (Suisse).

[Starner T. E. 1995]

Starner T. E. (1995). *Visual Recognition of American Sign Language Using Hidden Markov Models*. Master of Science in Media Arts and Sciences, MIT.

[Stokoe W. 1960]

Stokoe W. (1960). *Sign Language Structure: An Outline of the Visual Communication System of the American Deaf*. Studies in Linguistics, (University of Buffalo Press), Buffalo, NY.

[Sturman D. J. 1992]

Sturman D. J. (1992). *Whole-hand Input*. PhD thesis, MIT.

[Sturman D. J. et Zelter D. 1994]

Sturman D. J. et Zelter D. (1994). *A survey of Glove-based Input*. Computer Graphics and Applications, vol.14, pp.30-39.

[Takahashi T. et Kishino F. 1991]

Takahashi T. et Kishino F. (1991). *Hand Gesture Coding Based on Experiments Using a Hand Gesture Interface Device*. SIGCHI Bulletin, vol.23, pp.67-74.

[Tamura S. et Kawasaki S. 1988]

Tamura S. et Kawasaki S. (1988). *Recognition of Sign Language Motion Images*. Pattern Recognition, vol.21, pp.343-353.

[Thorisson K. R., Koons D. B. et al. 1992]

Thorisson K. R., Koons D. B. et Bolt R. A. (1992). *Multi-Modal Natural Dialogue*. Actes de CHI'92, pp.653-654.

[Torguet P., Rubio F. et al. 1995]

Torguet P., Rubio F. et Caubet R. (1995). *Atelier de sculture virtuelle multi-utilisateurs*. Actes de IHM'95, Cépaduès-éditions, pp.95-102, Toulouse.

[Touati M. 1983]

Touati M. (1983). *Structure complexe de la signation dans la langue des signes*. Autour de la langue des signes, (Université Paris V, UER de linguistique générale et appliquée), Paris, pp.17-25.

[Vamplew P. 1993]

Vamplew P. (1993). *The SLARTI Sign Language Recognition System: A Progress Report*. Communication personnelle, Adelaide (Australie).

[Zimmerman T. G., Lanier J. et al. 1987]

Zimmerman T. G., Lanier J., Blanchard C., Bryson S. et Harvill Y. (1987). *A Hand Gesture Interface Device*. Actes de CHI+GI'87, ACM, pp.189-192.[2]

[VPL 1989]

VPL (1989). *DataGlove Model 2 Users*. VPL - Research Inc.

ANNEXES

1.1. GROUPE DE TRAVAIL

(Extrait du rapport scientifique LIMSI 1994)

LE GESTE LIE A LA PAROLE

son apport pour un enrichissement des modèles cognitifs dans la communication humaine, appliqués au domaine de la communication homme-machine.

Responsable : Françoise FOREST

Objet Ce groupe de travail, soutenu par le pôle Paris-Sud du réseau Cognisciences, regroupe des chercheurs et étudiants de plusieurs organismes (universités Paris-V, Paris VIII, Paris-XI et Franche-Comté, ENS Fontenay-Saint-Cloud, les groupes L&C et CNV du LIMSI/CNRS, SERAC, INJS, ALSF, Acti-system).

Dans la suite de l'action incitative interne au LIMSI obtenue conjointement par M.F. Castaing et F. Forest il y a deux ans, les objets du groupe sont : (1) Approfondissement de nos connaissances sur le geste, (2) rencontre avec des collègues d'autres disciplines travaillant sur le geste de communication, (3) création et diffusion de documents concernant des résultats qui nous paraissent intéressants, soit en tant que résultats théoriques dans le domaine de la communication gestuelle, soit comme contribution au domaine de la communication homme-machine multimodale.

Contenu L'approche selon laquelle les travaux ont été orientés cette année structure le champ de recherche suivant deux axes. D'une part tout ce qui concerne le geste lui-même, sa description, sa représentation iconique, sa sémantique, les problèmes posés par sa capture, les différentes typologies, les rôles qu'il joue dans la communication, accompagné ou non de la parole… D'autre part tout ce qui concerne son apport théorique au domaine de la représentation interne profonde des connaissances, notamment par une démarcation d'avec les modèles traditionnels de représentation directement inspirés des langues linéaires (représentations logiques, réseaux sémantiques… plus généralement toutes les approches symboliques), son lien avec les images mentales, la suggestion de modèles informatiques incluant les notions de topologie, de continuité, de dynamique…

Situation Des rencontres et des échanges interdisciplinaires

Les origines diverses des participants au groupe ont motivé la tenue au LIMSI de plusieurs journées de travail, le 26 avril 1993, le 15 juin 1993, le 7 décembre 1993. Les différents participants du groupe y ont exposé les résultats auxquels ils étaient parvenus à partir d'un domaine restreint de gestes liés à l'expression du temps. A cette occasion, deux documents vidéo ont été réalisés : "le geste coverbal expression du temps", par G. Calbris, et "l'expression du temps dans la langue des signes", par C. Cuxac. L'ensemble des contributions devraient faire l'objet d'un document écrit que nous souhaitons publier dans les "Notes et documents" du LIMSI. Un corpus de gestes numérisés est en cours d'élaboration.

Une mise au point bibliographique sur les systèmes de codage du geste

Elle a été réalisée par Xavier Martin-Dupont dans le cadre de l'action incitative citée plus haut. Ce travail en cours de mise en forme donnera lieu à un document constitué d'une partie théorique et d'une partie pratique que nous souhaitons mettre à la disposition des collègues confrontés au problème de la capture et de la représentation informatiques de ces gestes.

La réalisation de documents de travail sur support approprié

L'aide du pôle Cognisciences Paris-Sud a permis d'aboutir au pressage d'un vidéodisque qui contient un corpus de gestes antérieurement réalisé par G. Calbris dans le cadre du CREDIF (ENS Fontenay-Saint-Cloud). Nous nous préoccupons actuellement de le rendre facilement accessible, afin de permettre aux chercheurs intéressés de le consulter facilement, et de bénéficier de l'organisation et du découpage théorique que G. Calbris a entrepris, avec l'aide de M. Hurault-Plantet, M.F. Castaing et de J. Ritter pour les aspects réalisation.

Participants Annelies Braffort, Rémi Brun, Geneviève Calbris, Marie-Françoise Castaing, Christophe Collet, Christian Cuxac, Françoise Forest, Sylvie Gibet, Martine Hurault-Plantet, Françoise Legault-Demare, Jimmy Leix, Xavier Martin-Dupont, Jacques Montredon, Jean Ritter, Jacob Sosnowski, Daniel Teil.

1.2. HISTOIRE DE LA LSF

Quelques points de repère :

17ème siècle :

Avant le 17ème : peu de témoignages (Platon - Aristote : quelqu'un qui ne parle pas ne peut pas raisonner).

Prêtres : l'éducation des sourds est possible - éducation oraliste.

18ème siècle :

- 1779 : premier livre écrit par un sourd, Pierre Desloges, sur la Langue des Signes et sa structure.
- 1712-1789 : Abbé de l'épée
  - Les gestes expriment la pensée humaine, comme une langue orale.
  - Création de la première école pour enfants sourds -> Institut St Jacques à Paris.
  - Invention des signes méthodiques (signes artificiels qui servent à calquer la grammaire du français). Ça fonctionne pour la dictée visuelle, mais ça n'a aucun sens pour les sourds.
- 1791 : L'assemblée nationale promulgue une loi permettant aux sourds de bénéficier des droits de l'homme.
- 1742 - 1822 : Abbé Sicard

  Système des signes méthodiques en pire.
- 1749-1834 : Bébian
  - Véritable éducation bilingue ;
  - Les élèves sourds deviennent professeurs pour enfants sourds ;
  - Les écoles se multiplient.
- 1816 : Laurent Clerc
  - Professeur à l'institut ;
  - Part aux États-Unis avec Gallaudet (américain) ;
  - Création de la première école pour enfants sourds aux États-Unis

    => influence de la LSF dans l'ASL.

Milieu 19ème siècle :

- Développement de la culture sourde ;
- Création d'associations ;
- Écrivains, peintres, poètes, conteurs…
- Premières études linguistiques ;

mais…

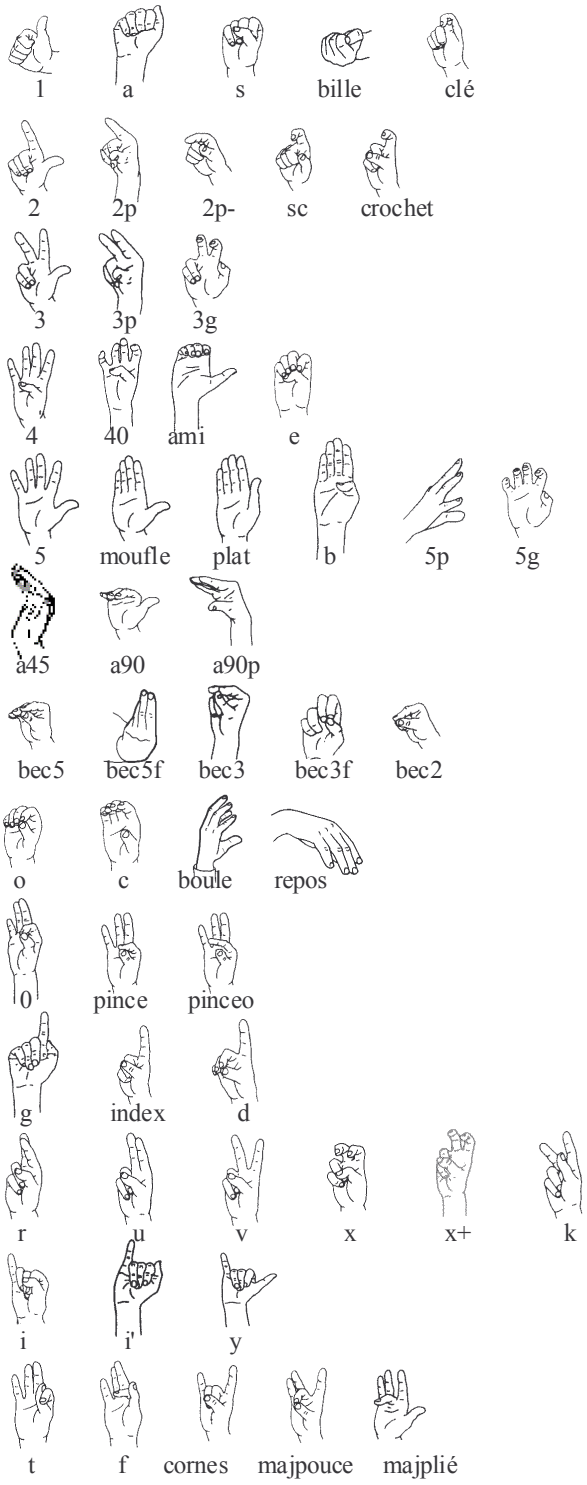
- Constantes querelles entre les tenants de l'éducation oraliste et ceux de l'éducation bilingue ;
- 1880 : Congrès international de Milan => Interdiction totale de la LSF ;

20ème siècle :

- 1968 : Droit à la parole donné aux minorités linguistiques ;
- 1975 : Journal hebdomadaire traduit en LSF sur la 2ème chaîne TV ;
- 1991 : Loi Fabius autorisant le choix de la langue pour les sourds ;
- 1992 : Emmanuelle Laborit, comédienne sourde, reçoit le Molière pour la pièce "Les enfants du silence" => porte-parole de la culture Sourde ;
- 1994 : 1ère émission TV en LSF et en français par et pour les sourds, sur la 5ème chaîne ;
- 1995 : Le comité d'éthique se prononce en faveur de l'apprentissage de la LSF en cas d'implant cochléaire sur les jeunes enfants.

2.1. CONFIGURATION

2.1.1. LES PRINCIPALES CONFIGURATIONS STATIQUES



La diversité et la répartition des configurations dépend du nombre de mains employées, ainsi que de ce rôle fonctionnel. Nous avons étudié la main dominante puis la main dominée.

2.1.2. CONFIGURATIONS DE LA MAIN DOMINANTE

Pour la main dominante, nous avons distingué trois cas : le cas général, dans lequel on considère tous les signes (à une et deux mains), le cas des signes à une main et le cas des signes à deux mains.

2.1.2.1. Cas général

Le cas général recouvre tous les signes, qu'ils soient réalisés avec une ou deux mains. Nous avons répertorié toutes les configurations rencontrées pour la main dominante. Nous avons dénombré **139** configurations différentes. Pour donner une idée de la répartition des signes parmi ces configurations, on peut indiquer que 58 configurations n'apparaissent qu'une fois dans le dictionnaire et seules 65 configurations apparaissent plus de deux fois. Nous listons ci-dessous les 20 configurations les plus rencontrées et leurs occurrences.

Nom	index	plat	moufle	s	5	1	v	a90	y	bec5
Occur.	122	78	71	69	59	52	52	40	38	36
Nom	clé	boule	pince	x	u	sc	c	crochet	majplié	5p/bec5
Occur.	35	31	30	30	24	23	22	22	20	18

2.1.2.2. Cas des gestes à une main

Les signes qui sont réalisés avec une seule main sont au nombre de 471, soit 37,5 % des 1257 signes. On observe dans ce cas **92** configurations différentes, dont 42 n'apparaissent qu'une fois et 41 apparaissent plus de deux fois. Nous listons ci-dessous les 20 configurations les plus rencontrées et leurs occurrences.

Nom	index	plat	1	v	y	s	moufle	5	bec 5	a90
Occur.	53	32	25	23	20	18	17	16	13	12
Nom	clé	boule	pince	x	2p/bec2	c	majplié	crochet	0	su
Occur.	12	12	10	10	9	9	9	8	7	7

2.1.2.3. Cas des gestes à deux mains

Les signes qui sont réalisés avec deux mains sont au nombre de 786, soit 62,5 % des signes. On observe dans ce cas **107** configurations différentes, dont 43 n'apparaissent qu'une fois et 51 apparaissent plus de deux fois. Nous listons ci-dessous les 20 configurations les plus rencontrées et leurs occurrences.

Nom	index	moufle	s	plat	5	v	a90	1	bec 5	clé
Occur.	69	54	51	46	43	29	28	27	23	23

Nom	pince	x	boule	y	sc	u	5p/bec5	crochet	c	majplié
Occur.	20	20	19	18	17	17	14	14	13	11

2.1.3. CONFIGURATIONS DE LA MAIN DOMINEE

Pour la main dominée, nous avons distingué trois cas : le cas général, le cas où les configurations des deux mains sont identiques et le cas où elles sont différentes.

2.1.3.1. Cas général

Le cas général comporte tous les cas où les deux mains sont utilisées. On observe **99** configurations différentes, dont 45 n'apparaissent qu'une fois et 43, plus de deux fois. Nous listons ci-dessous les 20 configurations les plus rencontrées et leurs occurrences.

Nom	moufle	plat	s	index	5	a90	1	bec 5	boule	v
Occur.	114	86	77	58	49	28	19	19	18	17
Nom	clé	sc	u	c	x	y	pince	crochet	0	s/5
Occur.	16	16	15	13	12	11	11	10	9	8

2.1.3.2. Cas où la configuration des mains est identique

Dans plus de 75 % des cas, la configuration de la main dominée est identique à celle de la main dominante. Dans ce cas, on observe **93** configurations différentes, dont 40 n'apparaissent qu'une fois et 40, plus de deux fois.

Nom	index	moufle	5	plat	s	a90	1	boule	bec 5	v
Occur.	48	47	44	43	42	24	19	18	17	17
Nom	clé	sc	u	x	y	c	crochet	pince	0	s/5
Occur.	16	16	13	12	11	10	10	10	8	8

2.1.3.3. Cas où la configuration des mains est différente

Pour les 25 % restants, on n'observe que **21** configurations différentes, dont 10 (toutes statiques) apparaissent plus de deux fois et 9 (dont 3 dynamiques), une seule fois. Voici les 4 configurations les plus rencontrées dans ce cas.

Nom	moufle	plat	s	index
Occur.	67	43	35	10

2.2. MOUVEMENT

2.2.1 PRIMITIVE DROITE

A l'aide du repère décrit au Paragraphe 2.3.2.1, nous avons étudié la distribution des différents signes de la classe droite. Nous avons obtenu les résultats présentés dans le tableau suivant.

Type	Occurrence	%
[X]	180	33,4
[Y]	127	23,6
[Z]	137	25,4
[X,Y]	42	7,8
[Y,Z]	17	3,2
[X,Z]	21	3,9
[V]	5	0,9
[X,Y,Z]	10	1,9

Le tableau suivant précise la répartition des signes sur les différentes directions possibles pour les sous-classes les plus fréquentes de la classe droite. Pour les plans, deux symboles + ou - désignent respectivement le premier et le deuxième axe. On distingue ainsi quatre parties dans chaque plan.

Direction	[X]	[Y]	[Z]
+	113	34	73
-	67	103	54
Direction	**[XY]**	**[YZ]**	**[XZ]**
+ +	0	1	11
+ -	15	2	15
- +	6	13	14
- -	0	1	2

2.2.2 PRIMITIVE ARC

La distribution des différents signes de la classe arc est présentée dans le tableau suivant.

Type	Occurrence	%
[X,Y]	141	43
[Y,Z]	81	24,7
[X,Z]	66	20,1
[V]	24	7,3
[X,Y,Z]	16	4,9

Le tableau suivant précise, pour chaque plan du repère et pour chaque sens, les huit dénominations utilisées pour distinguer les arcs.

Zones	[X,Y] +	[X,Y] -	[Y,Z] +	[Y,Z] -	[X,Z] +	[X,Z] -
1	avant	arrière	droite	gauche	droite	gauche
2	avant-bas	haut-arrière	droite-bas	haut-gauche	droite-arrière	avant-gauche
3	bas	haut	bas	haut	arrière	avant
4	bas-arrière	avant-haut	bas-arrière	droite-haut	arrière-gauche	droite-avant
5	arrière	avant	arrière	droite	gauche	droite
6	arrière-haut	bas-avant	arrière-haut	bas-droite	gauche-avant	arrière-droite
7	haut	bas	haut	bas	avant	arrière
8	haut-avant	arrière-bas	haut-avant	gauche-bas	avant-droite	gauche-arrière

Le tableau suivant précise la répartition des arcs dans les différentes sous-classes ainsi distinguées :

Zone	[XY] +	[XY] -	[YZ] +	[YZ] -	[XZ] +	[XZ] -
1	21	5	6	9	2	15
2	36	8	1	0	0	10
3	5	5	4	0	3	1
4	1	2	4	4	0	2
5	4	15	10	2	0	3
6	3	18	3	5	2	0
7	7	5	3	6	10	0
8	4	2	13	11	12	6

2.2.3 PRIMITIVE CERCLE

La distribution des différents signes de la classe cercle est présentée dans le tableau suivant. Il précise la répartition des signes sur les différentes directions possibles pour les sous-classes les plus fréquentes de la classe cercle. On note +, les cercles effectués dans le sens des aiguilles d'une montre et - le sens inverse, en se plaçant selon les mêmes points de vue que pour les arcs.

Direction	[XY]	[YZ]	[XZ]
+	37	8	7
-	17	37	20

2.3. ORIENTATION

Le tableau suivant contient les 19 orientations les plus rencontrées pour la main droite (au moins 10 occurrences) :

Axe main	Paume	Occurrences	Pourcentage
haut	gauche	112	8,9
gauche	arrière	89	7,1
haut	avant	76	6,1
avant	gauche	69	5,5
haut	arrière	63	5
avant	bas	56	4,5
avant-gauche	bas	40	3,2
gauche	bas	37	2,9
avant	haut	27	2,2
haut/avant	gauche	24	1,9
gauche-haut	arrière	23	1,8
gauche-haut	bas-gauche	19	1,5
avant-gauche	gauche-arrière	13	1
haut/avant	avant/bas	12	1
gauche	haut	11	0,9
haut/gauche	gauche/bas	11	0,9
avant/haut	gauche	10	0,8
gauche	arrière/bas	10	0,8
gauche	bas/arrière	10	0,8

Le tableau suivant contient, pour chaque primitive de mouvement, les proportions d'orientations statiques et dynamiques :

Orientation	droite	arc	cercle	statique
statique	87,4 %	20,1 %	93,4 %	38 %
dynamique	12,6 %	79,9 %	6,6 %	62 %

Le tableau suivant contient la proportion de mouvements du poignet pour chaque primitive de mouvement :

Combinaison	droite	arc	cercle	statique
aucune	92,4 %	42,7 %	90,5 %	33 %
rotation et/ou flexion	7,6 %	57,3 %	9,5 %	67 %

2.4. EMPLACEMENT

Le tableau suivant contient les 14 emplacements les plus fréquents pour la main droite :

Zone	Pourcentage
devant torse	61,1
devant tête	6,3
torse	4,1
bouche	3,7
front	3,4
menton	3,1
coté tête	2,8
cou	1,8
joue	1,5
bras	1,1
ventre	1
yeux	1
torse gauche	0,9
nez	0,8

2.5. AUTRES INFORMATIONS

Le tableau suivant contient les proportions de répétition pour chaque primitive de mouvement. Ces répétitions concernent soit la configuration, soit le mouvement, soit l'orientation.

Répétition	statique	droite	arc	cercle
aucune	31,6 %	51,9 %	80,8 %	8 %
une	67 %	46,6 %	18,9 %	92 %
retour	1,4 %	1,5 %	3,3 %	0 %

Le tableau suivant indique les rapports entre les mouvements des deux mains et leur configuration, pour chaque primitive de mouvement :

Mouvement	Mains	statique	droite	arc	cercle
1 seule main	1	43,1	34,7	34,1	43,8
2 - identique	2 - identique	17,4	20,6	13,4	5,1
2 - différent	2 - identique	6	8,9	15,5	10,9
2 - différent	2 - différent	13,8	16	14,9	10,9
2 - opposé	2 - identique	17,9	17,4	21,3	13,1
2 - décalé	2 - identique	1,8	2,4	0,6	16,1

Les tableaux suivants contiennent le pourcentage de signes sans contact, les pourcentages de signes avec contact en fonction de la partie du corps concernée, puis les pourcentages de signes avec contact en fonction du moment pendant lequel le contact a lieu :

Critère	Occurrence	Pourcentage
pas de contact	576	45,8

Critère	Occurrence	Pourcentage
2 mains	395	31,4
main visage	163	13
main corps	96	7,6
main bras	27	2,2

Critère	Occurrence	Pourcentage
fin du signe	231	18,4
début du signe	192	15,3
tout le temps	158	12,6
milieu du signe	74	5,9
autre	26	2

3. RECONNAISSANCE

Cette annexe contient un tableau récapitulatif de l'état de l'art en reconnaissance de gestes. Il est classé par ordre alphabétique de nom d'auteur. Il comporte 9 colonnes qui contiennent :

- le premier nom d'auteur des articles ou rapports concernant l'étude,
- l'année de l'article,
- le système de capture utilisé :
  - DataGlove (VPL)
  - CyberGlove (Virtual Technology)
  - PowerGlove (Mattel)
  - Gant X (Nom non fourni)
  - Caméra
- le type de représentation:
  - Prototypes (Prot.), éventuellement associés à une approche statistique (stat.)
  - Discrétisation de l'espace en zones (Zones)
  - Caractéristiques cinématiques et géométriques (Carac.)
- le processus de décision:
  - Linguistique
  - Comparaison de prototypes
  - Arbre binaire de décision (ABD)
  - Réseaux connexionnistes (NN) de type Perceptron, KFM ou récurrent
  - Comparaison dynamique (Comp. dyn.)
  - Modèles de Markov cachés (HMM)
- la technique de segmentation en cas de gestes connectés ou enchaînés :
  - Pauses
  - Fenêtre
  - Configuration
  - Caractéristiques cinématiques et géométriques
- la constitution du corpus :
  - Postures : configuration, éventuellement position et orientation, mais pas de mouvement
  - Gestes : configuration, position, orientation et mouvement
  - Phrases : séquences de gestes enchaînés
  - Mouvements : pas de configuration, mais position, orientation et mouvement
- le type de corpus :
  - Isolés
  - Connectés
  - Enchaînés
- Les taux de reconnaissance, lorsqu'ils sont fournis.

Nom	Année	Capteur	Représenteur	Décideur	Segmentation	Postures Gestes PHrases Mouvement	Isolés Connectés Enchainés	Taux
Bellalem	95	DataGlove	Carac.	Linguistique		G	I	
Boehm	94	CyberGlove	Prototypes	KFM		10 G	I	
			Carac. + KFM	NN récurrent		G	I	
Bordegoni	93	DataGlove	Zones	Comparaison	Config	6 G	E	
Braffort	92	DataGlove	Zones	ABD		10 P	E	
	92		Carac.	Comparaison	Config	10 G	E	
	96		Prot. + Carac.	HMM	HMM	44 PH de 4 G	E	96 et 92
Cagin	93	DataGlove	Prot. + stat	ABD	Pauses	25 P	C	99
			Prototypes	Comp. dyn.		14 M	I	99,1
			Prototypes	Comp. dyn.	Comp. dyn.	14 M	C	93,9
			les 2	les 2	les 2	14 G	C	87
Collet	93	DataGlove	Prototypes + compression	NN Perceptron	C.géo + manuel	28 G	I + E	90
Dasilva	91	DataGlove	Prot. + stat.	Comparaison	Fenêtre	5 P	E	
			Prototypes	Comp. dyn.	Pauses	10 G	E	
Fels	90	DataGlove	Prot. + Carac.	NN Percep. //	NN	66 P + 6 G	C	94
	94	CyberGlove	Prot. + Carac.	NN Percep. //	Pédale		C	
Gourley	94	CyberGlove	Prototypes	NN Perceptron		26 P	I	95
Hand	94	PowerGlove	Prototypes ?	Linguistique		8P + 6 G	I	15 -> 80
Harling	93	PowerGlove	Prototypes	NN Perceptron		5 P	I	96
						3 G	I	91,4
Kadous	95	PowerGlove	Carac.	Linguistique		95 G	I	50
			Carac.	Comparaison		95 G	I	80
Kramer	89	CyberGlove	Prototypes	NN Perceptron		26 P	I	
Liang	95	DataGlove	Zones	Comparaison	Fenêtre/Pauses	26 P	E	

Nom	Année	Capteur	Représenteur	Décideur	Segmentation	Postures Gestes PHrases Mouvement	Isolés Connectés Enchainés	Taux
Messing	94	CyberGlove	Prototypes	NN Perceptrons en cascade		26 P	I	96,5
Murakami	91	DataGlove	Prototypes Prot. + Carac.	NN Perceptron NN récurrent	Seuil Config+Fenêtre	42 P 10 G	I E	98 96
Nam	95	DataGlove	Prototypes + compression	HMM	Caractéristiques + HMM	10 G	C	80
Newby	93	DataGlove	Prototypes	Comparaison		36 P	I	
Revesz	93	CyberGlove	Zones	Comparaison		365 P	I	96
Sagawa	92	DataGlove	Carac.	Comp. dyn.	Comp. dyn.	5P + 17 G	E	97,3
Searles	93	Polhemus	Zones Carac.	Comparaison Comparaison	Fenêtre Carac.	5P + 6G 5P + 6G	E E	93 et 80 82
Sparrel	93	DataGlove	Carac	ad hoc	Pauses + parole (Multimodal)	G	E	
Starner	95	Caméra	Prot. + Carac.	HMM	HMM	99PH (40G) 1 struct. gram.	E	95 97
Sturman	92	DataGlove	Carac.	ad hoc	Fenêtre		E	
Takahashi	91	DataGlove	Zones	Comparaison	Fenêtre/Pauses	46 P	E	65
Tamura	88	Caméra	Carac.	Décision hiérar.	Positions clé	20 G	C	45
Vamplew	93	CyberGlove	Prototypes	NN Perceptron		~ 30 P	I	
Watson	95	Gant X	Carac.	Comparaison		12 P	I	simulé
Zimmerman	87	DataGlove	Zones	Comparaison	Fenêtre	10 P	E	

TABLE DES MATIERES

Introduction **1**

1. La communication gestuelle **5**

1.1. Positionnement du problème *6*

1.1.1. Les trois fonctions du geste humain 6
1.1.2. Les domaines d'application en CHM 7
1.1.3. Les systèmes de capture de gestes 9
1.1.4. Thème de recherche étudié 11

1.2. Le geste de commande *12*

1.2.1. Les interfaces gestuelles 12
1.2.2. Un modèle d'interaction gestuelle 20
1.2.3. Bilan 25

1.3. Le geste co-verbal *26*

1.3.1. Fonctions des gestes co-verbaux 26
1.3.2. Informations spatiales multimodales 30
1.3.3. Informations temporelles 36
1.3.4. Conclusion 38

1.4. Le geste de la langue des signes *39*

1.4.1. Les différences entre les langues orales et les langues gestuelles 39
1.4.2. Les Paramètres 40
1.4.3. Phonologie 46

1.5. Conclusion *48*

2. Étude des paramètres de la LSF **49**

2.1. Introduction *50*

2.2. Définition des paramètres *51*

2.3. Les quatre paramètres *53*

2.3.1. Configuration 53
2.3.2. Mouvement 68
2.3.3. Orientation 79
2.3.4. Emplacement 84

2.4. Autres informations *89*

2.5. La base de données *91*

2.5.1. Structure complète 91
2.5.2. Exemple de description 93

2.6. Conclusion *94*

3. La reconnaissance de gestes 95

3.1. État de l'art 96

3.1.1. Terminologie 96
3.1.2. Segmentation 99
3.1.3. La représentation 101
3.1.4. La décision 106
3.1.5. Applications et corpus 121
3.1.6. Conclusion 124

3.2. Outils utilisés et développés 126

3.2.1. Le système de capture de gestes et ses limitations 126
3.2.2. VAG 2 : Visualisation et Analyse de Gestes 132
3.2.3. TePa : Outil de Test des Paramètres 134
3.2.4. Autres outils 135

3.3. Expérimentations réalisées 136

3.3.1. Le module de reconnaissance 136
3.3.2. Le corpus 146
3.3.3. Mise en oeuvre des HMM 154
3.3.4. Premiers résultats 160

3.4. Conclusion 162

4. Vers un système de compréhension 163

4.1. Modélisation de la scène de narration 164

4.1.1. Représentation des entités 164
4.1.2. Représentation de la scène de narration 168

4.2. Architecture du système ARGo 170

4.2.1. Analyseur 171
4.2.2. Intégrateur 175
4.2.3. Les bases de données 177
4.2.4. La scène virtuelle 177

4.3. Fonctionnement : un exemple 178

4.4. Évaluation 185

4.5. Applications 186

4.6. Conclusion 191

Conclusion et perspectives 193

Bibliographie 197

Annexes.....211

1.1. Groupe de travail 211

1.2. Histoire de la LSF 212

2.1. Configuration 213

2.2. Mouvement 217

2.3. Orientation 219

2.4. Emplacement 220

2.5. Autres informations 221

3. Reconnaissance 222

Table des matières 225